DEVELOPMENT OF ATOM-CENTERED POTENTIALS FOR EFFICIENT AND ACCURATE ELECTRONIC STRUCTURE MODELING OF LARGE MOLECULAR SYSTEMS

by

Viki Kumar Prasad

B.Sc., St. Xavier's College, India, 2014

M.Sc., Indian Institute of Technology Kharagpur, India, 2016

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The College of Graduate Studies

(Chemistry)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

December 2021

The following individuals certify that they have read, and recommend to the College of Graduate Studies for acceptance, a thesis/dissertation entitled:

DEVELOPMENT OF ATOM-CENTERED POTENTIALS FOR EFFICIENT AND ACCURATE ELECTRONIC STRUCTURE MODELING OF LARGE MOLECULAR SYSTEMS

submitted by Viki Kumar Prasad in partial fulfillment of the requirements of

the degree of Doctor of Philosophy.

Examining Committee:

Prof. Gino A. DiLabio, Irving K. Barber Faculty of Science

Supervisor

Dr. Alberto Otero-de-la-Roza, University of Oviedo

Co-supervisor

Dr. Isaac T.S. Li, Irving K. Barber Faculty of Science

Supervisory Committee Member

Dr. David Jack, Irving K. Barber Faculty of Science

Supervisory Committee Member

Dr. Joshua Brinkerhoff, School of Engineering

University Examiner

Prof. Erin Johnson, Dalhousie University

External Examiner

Abstract

Accurate quantum mechanical (QM) modeling of large molecular systems is computationally challenging due to the dramatic increase in the demand for computational resources with increasing system size. To tackle this problem, atom-centered potentials (ACPs) were developed to mitigate the errors of Hartree–Fock (HF) and density-functional theory (DFT) methods, particularly when used with small basis sets. The objective behind developing ACPs for such methods was to improve their accuracy in predicting various molecular properties without impacting their low computational cost. ACPs are optimized one-electron Gaussian-type functions that share the same mathematical form as generally used effective-core potentials, except they do not replace any electrons, making them easily usable with many quantum chemistry software packages. Besides, ACPs allow for a convenient means to simultaneously correct the absence of correlation (or deficiencies in exchange-correlation density functionals), basis set incompleteness, and other shortcomings in HF or DFT methods with small basis sets. The overall research conducted for this dissertation demonstrates the gradual transition from the development of proof-of-concept ACPs to final ACPs with more general applicability. In particular, the final ACPs are presented for ten elements in the first and second rows (H, B, C, N, O, F, Si, P, S, Cl), extending the applicability to various organic and biochemical molecules. The ACP-corrected methods have been shown to predict the target molecular properties with slightly less accuracy than very computationally expensive QM methods but at a much lower computational cost. It is anticipated that the methods presented in this dissertation will assist in applications such as supramolecular host-guest complexation, enzymatic catalysis, drug-target binding, protein folding, and others. This dissertation also contributes towards filling the gap in the literature regarding benchmark data sets by presenting new diverse data sets of molecular properties such as polypeptide conformational energies (PEPCONF), bond separation energies (BSE49), barrier height energies (BH9), and reaction energies (BH9-RE). These data sets have been generated using a significant amount of manual and computational effort to address the need for reference data in the ACP development process and other applications.

Lay summary

The utility of computer modeling as a predictive and explicative tool, or complementary to experiments, continues to help advance (bio)chemistry. It has impacts in finding potential drug candidates, discovering new materials with various applications, investigating microscopic chemical processes, etc. Computational chemists rely on mathematical methods incorporated into computer programs to describe the behaviour of electrons in atoms and molecules and predict their structure and properties. However, the computational power required by more accurate methods increases rapidly with the system size. Therefore, continual efforts are being made to improve the balance between the efficiency and accuracy of computational chemistry methods. This dissertation highlights such an effort to develop new computationally inexpensive methods that enable the modeling of large systems with reduced use of computational power but with better accuracy. The methods developed are valuable for those interested in (bio)chemical structure-function relationships, *a priori* structure determination, and *in silico* design and discovery of new materials.

Preface

The research chapters (Chapters 3–9) in this dissertation have been prepared by including a collection of published articles and submitted manuscripts. Supplementary data accompanying these chapters have been deposited to the figshare repository and is openly available at the following URL/DOI: https://doi.org/10.6084/m9.figshare.16912201. All research studies were co-supervised by Gino A. DiLabio and Alberto Otero-de-la-Roza. The details about the materials forming the research chapters are provided as follows:

Chapter 3

"*Atom-centered potentials with dispersion-corrected minimal-basis-set Hartree–Fock: An efficient and accurate computational approach for large molecular systems*" by Viki Kumar Prasad, Alberto Otero-de-la-Roza, and Gino A. DiLabio

Received: Nov. 15, 2017; *Published (online): Jan. 16, 2018; Published (issue): Feb. 13, 2018*

Reprinted (adapted) with permission from *Journal of Chemical Theory and Computation* 2018, 14(2), 726–738. © Copyright 2018 American Chemical Society. (DOI: 10.1021/acs.jctc.7b01158)

Copyright information: https://pubs.acs.org/page/copyright/permissions_journals.html

Supplementary material: The supporting information is available free of charge on the ACS Publications website at https://doi.org/10.1021/acs.jctc.7b01158. Some parts of the supporting information are also provided in Appendix 1.

Author contribution: The research study was conceptualized and designed by Gino A. DiLabio and Alberto Otero-de-la-Roza. Viki Kumar Prasad executed the various parts of the research work including development of methodology, performance of calculations, analysis of data, interpretation of data, validation of data, curation of data, and preparing the first draft of the manuscript.

Chapter 4

"*PEPCONF, a diverse data set of peptide conformational energies*" by Viki Kumar Prasad, Alberto Otero-de-la-Roza, and Gino A. DiLabio

Received: Oct. 15, 2018; *Accepted: Nov. 30, 2018; Published: Jan. 22, 2019*

Reprinted (adapted) with permission from *Scientific Data* 2019, 6, 180310. © Copyright 2019 Springer Nature Limited. (DOI: 10.1038/sdata.2018.310)

Copyright information: https://www.nature.com/sdata/policies/editorial-and-publishing-policies

Supplementary material: The supporting information is available free of charge on the Nature Research website at https://doi.org/10.1038/sdata.2018.310. The PEPCONF dataset is available on the Figshare repository website at https://doi.org/10.6084/m9.figshare.7185194.v2. Some parts of the supporting information are also provided in Appendix 2.

Author contribution: The research study was conceptualized by Viki Kumar Prasad and designed by Gino A. DiLabio. Viki Kumar Prasad executed the various parts of the research work including generation of data set, performance of calculations, analysis of data, interpretation of data, validation of data, curation of data, and preparing the first draft of the manuscript.

Chapter 5

"*BSE49, a diverse, high-quality benchmark dataset of separation energies of chemical bonds*" by Viki Kumar Prasad, M. Hossein Khalilian, Alberto Otero-de-la-Roza, and Gino A. DiLabio

Submitted: May 6, 2021; Accepted: Nov. 1, 2021; Published: Nov. 23, 2021

Reprinted (adapted) with permission from *Scientific Data* 2021, 8, 300. Copyright © 2021 Springer Nature Limited. (DOI: 10.1038/s41597-021-01088-2)

Copyright information: https://www.nature.com/sdata/policies/editorial-and-publishing-policies

Supplementary material: The BSE49 dataset is available on the Figshare repository website at https://doi.org/10.6084/m9.figshare.14544060.v1.

Author contribution: The research study was conceptualized and designed by Viki Kumar Prasad. Viki Kumar Prasad also executed the various parts of the research work including generation of data set, performance of calculations, analysis of data, interpretation of data, validation of data, curation of data, and preparing the first draft of the manuscript. M. Hossein Khalilian assisted in the curation and validation of data.

Chapter 6

"*BH9, a new comprehensive benchmark dataset for barrier heights and reaction energies: Assessment of density functional approximations and basis set incompleteness potentials*" by Viki Kumar Prasad, Zhipeng Pei, Simon Edelmann, Alberto Otero-de-la-Roza, and Gino A. DiLabio

Received: Jul. 10, 2021; Published (online): Dec. 15, 2021; In *Journal of Chemical Theory and Computation* 2021. Copyright © 2021 American Chemical Society. (DOI: 10.1021/acs.jctc.1c00694)

Copyright information: https://pubs.acs.org/page/copyright/permissions_journals.html

Supplementary material: Some parts of the supporting information are provided in Appendix 3.

Author contribution: The research study was conceptualized and designed by Viki Kumar Prasad. Viki Kumar Prasad also executed the various parts of the research work including generation of data set, performance of calculations, analysis of data, interpretation of data, validation of data, curation of data, and preparing the first draft of the manuscript. Zhipeng Pei assisted in the curation and validation of data as well as helped with trouble-shooting some failed calculations. Simon Edelmann assisted in construction of geometries and location of initial transition states of some reactions in the data set. Alberto Otero-de-la-Roza ran all the necessary calculations for the benchmarking aspect of the work.

Chapter 7

"*Performance of small basis set Hartree–Fock methods for modeling non-covalent interactions*" by Viki Kumar Prasad, Alberto Otero-de-la-Roza, and Gino A. DiLabio

An invited article for the issue *Focus on Van der Waals Interactions* (https://iopscience.iop.org/journal/2516-1075/page/Focus_on_Van_der_Waals_Interactions)

Received: Jul. 8, 2021; *Accepted: Sep. 1, 2021; Published: Sep. 20, 2021*

Reprinted (adapted) with permission from *Electronic Structure* 2021, 3(3), 034007. © Copyright 2021 Institute of Physics Publishing Ltd. (DOI: 10.1088/2516-1075/ac22b8)

Copyright information: https://publishingsupport.iopscience.iop.org/reusing-iop-published-content/

Supplementary material: The supporting information is available free of charge on the Figshare repository website at https://doi.org/10.6084/m9.figshare.14933016. Some parts of the supporting information are also provided in Appendix 4.

Author contribution: The research study was conceptualized and designed by Gino A. DiLabio and Alberto Otero-de-la-Roza. Viki Kumar Prasad executed the various parts of the research work including development of methodology, performance of calculations, analysis of data, interpretation of data, validation of data, curation of data, and preparing the first draft of the manuscript.

Chapter 8

"*Fast and accurate quantum mechanical modeling of large molecular systems using small basis set Hartree–Fock methods corrected with atom-centered potentials*" by Viki Kumar Prasad, Alberto Otero-de-la-Roza, and Gino A. DiLabio

Submitted: Nov. 8, 2021; *Manuscript ID: ct-2021-01128c;* Submitted to *Journal of Chemical Theory and Computation* 2021. Copyright © 2021 American Chemical Society.

Copyright information: https://pubs.acs.org/page/copyright/permissions_journals.html

Supplementary material: Some parts of the supporting information are provided in Appendix 5.

Author contribution: The research study was conceptualized by Gino A. DiLabio and Alberto Otero-de-la-Roza. All authors participated equally in the design of the research. Viki Kumar Prasad executed the various parts of the research work including performance of calculations, analysis of data, interpretation of data, validation of data, curation of data, and preparing the first draft of the manuscript.

Chapter 9

"*Small basis set density-functional theory methods corrected with atom-centered potentials*" by Viki Kumar Prasad, Alberto Otero-de-la-Roza, and Gino A. DiLabio

To be submitted to *Journal of Chemical Theory and Computation* 2021. Copyright © 2021 American Chemical Society.

Copyright information: https://pubs.acs.org/page/copyright/permissions_journals.html

Supplementary material: Some parts of the supporting information are provided in Appendix 6.

Author contribution: The research study was conceptualized by Gino A. DiLabio and Alberto Otero-de-la-Roza. All authors participated equally in the design of the research. Viki Kumar Prasad executed the various parts of the research work including development of methodology, performance of calculations, analysis of data, interpretation of data, validation of data, curation of data, and preparing the first draft of the manuscript.

Table of Contents

List of Tables

List of Figures

Acknowledgements

The completion of this dissertation was made possible by the help and support of many people to whom I owe my sincerest appreciation.

Foremost, I would like to express my heartful gratitude to my advisor Prof. Gino DiLabio for noticing my potential and giving me the wonderful opportunity to conduct my PhD studies under his supervision. I want to thank him for continuously supporting me throughout my PhD and allowing me to work on many exciting projects over these years. I also acknowledge him for his remarkable mentorship at every step of my research and for providing valuable words of wisdom to succeed in the degree during challenging times. More importantly, my sincere thanks to Prof. DiLabio for being like a fatherly and friendly figure in my life to whom I look up to for inspiration. Thank you again for enriching my life with valuable skill sets and providing scientific nourishment to an aspiring researcher who traveled 11066 km away from what he calls home in pursuit of his passion for chemistry.

I want to mention my sincerest thanks to Dr. Alberto Otero-de-la-Roza for being an incredible co-supervisor throughout these years. The research conducted in this dissertation would not have been possible without his help and guidance. I thank him for his continuous support and encouragement. I also acknowledge him for his immense patience in dealing with my numerous questions during our research discussions. He has helped me grow to my fullest potential by improving my skills and instilling self-confidence. The valuable knowledge that I have acquired from him will last me till the end of my life.

My sincere thanks to Dr. Isaac Li and Dr. David Jack for their patience and for serving in my PhD committee all these years. I would like to recognize them for their valuable feedback and insights during my PhD studies, especially in helping me develop scientific inquiry and critical thinking skills. My special thanks to Dr. Erin Johnson and Dr. Joshua Brinkerhoff for agreeing to be the external examiner and university examiner, respectively. I express my thanks to Dr. Frederic Menard and Dr. Kevin Smith for serving in the departmental role of graduate student faculty advisor during my PhD at UBCO. My heartiest thanks to Dr. Paul Shipley for his help with fast-tracking the amendments to the university policies around thesis preparation.

I acknowledge the Natural Sciences and Engineering Research Council of Canada, Canadian Foundation for Innovation, B.C. Knowledge Development Fund, and the University of British Columbia for financial support. I would also like to acknowledge WestGrid, Compute Canada, and the University of

British Columbia's advanced research computing services, as their generous computational resource support enabled the research presented in this dissertation.

I owe my sincere thanks to the past and present members of the DiLabio research group for their help and constant support, especially Dr. Alireza Sadeghifar, M. Hossein Khalilian, Eduardo Romero-Montalvo, Dr. Emilian Tuca, Zhipeng Pei, An Goto, and Jeffrey van Santen.

I would also like to thank my various friends who welcomed me into their life during my PhD studies in Canada, particularly Dr. Hitendra Kumar, Vikas Kumar, and Nitin. A special mention to Daniel Jones and Dave Hallonquist for kindly providing a shelter in Canada that allowed a productive environment to thrive during my degree.

I thank my dearest friends located in India for their support from thousands of miles away that kept me motivated throughout these years, particularly Aritra Kar and Sailaja Agarwal. I would also like to thank Rahul Goswami, Kamini Agarwal, Aurea Dutta, Sanjukta Saha, Sagnik Bhattacharjee, Mofijul Mollah, Md. Raja Sk., Anurima De, Anibrata Biswas, and Dr. Konoya Das. My school classmates, Sumantika Banerjee and Pritam Dutta, you will also be in my fond memories and dearly missed.

Special thanks are owed to my beloved parents (Vijay Prasad and Premshila Prasad) and my brother (Sidhant Prasad) for their endless love and support. I also thank my cousins and my various relatives for their kindness. My generous love and gratefulness to Denise Kidd and Joseph Brien. Thank you for welcoming me into your family; I cherish every moment spent with you both. A special mention to the feathered and furry friends who had filled and still fill my life with joy: Missy, Maple, and Ion.

Last but not least, my gratitude to my partner in crime, Chelsey Brien, for supporting me during the challenging times and helping me never give up on my dreams. Thank you for believing in my potential and helping me be a better person every single day. You are an inspiration to many, including me.

Finally, I would like to mention my gratefulness, in no specific order, to Dr. Tamara Freeman, Dr. Jennifer Robison, Michele Cannon, Dr. Eric Dennis, Ieva Zigg, Susan Black, Wade Klaver, the staff of the chemistry department, and the staff of the College of Graduate Studies for all their kind help and support during my PhD studies at UBCO.

Dedication

यह शोध प्रबंध मेरी मम्मी, पापा और भाई को समर्पित है |

आप सभी इन चुनौतीपूर्ण वर्षों में निरंतर प्रेरणा के स्रोत रहे हैं।

This dissertation is dedicated to my mummy, papa, and brother.

All of you have been a constant source of motivation in these challenging years.

Part I

Chapter 1

Introduction

1.1 Research Background

Developments in quantum mechanics (QM) and modern computer technology have enabled tremendous advances in the field of computational chemistry. The widespread use has allowed the investigation of a wide range of chemical and biological problems.[1–3] Quantum chemistry methods are useful for initial exploration in experimental studies, validation of experimental data, and chemical research of topics where experiments are impractical, expensive, or inconvenient. They can be applied to determine the optimal three-dimensional arrangement of molecules as well as to predict, interpret, and rationalize various chemical properties such as non-covalent interactions, bond strengths, reaction energies, reaction rates, spectroscopic constants, and more. The success of computational chemistry has been demonstrated by the Nobel prize awards of 1998 and 2013. The 1998 award was granted to Walter Kohn and John A. Pople for their pioneering contributions to the development of computational methods in quantum chemistry.[5] The 2013 award was presented to Martin Karplus, Michael Levitt, and Arieh Warshel for developing new computationally inexpensive techniques applicable to complex (bio)chemical systems.[6]

An example that illustrates the explicative and predictive capabilities of quantum chemistry can be seen in Reference 4, where using a theoretical approach, the authors showed a remarkable agreement between the predicted rotational vibration spectrum and dissociation energy of the F_2 molecule and experiment, which had been measured earlier using high-resolution electronic spectroscopy. The mean absolute deviation between the theoretical and experimental spectrums was only about 5 cm^{-1}. At the same time, the calculated dissociation energy with respect to the lowest vibrational energy was shown to be within 30 cm^{-1} of the experimental value.

The ability to incorporate electronic effects in chemical modeling is an important feature of quantum mechanical methods. This allows such methods to model the electronic structure of molecules and obtain the related chemical properties, particularly those in which electronic rearrangements or the formation and breaking of chemical bonds occur. In this context, the focus of this dissertation excludes other computational methods beyond the atomistic resolution, like coarse-grained[7] methods, or that do not allow modeling of electronic effects, like classical force fields[8].

The starting point of any quantum mechanical calculation is the definition of molecular information such as overall charge, spin multiplicity, atom composition, and three-dimensional structure. Furthermore, a choice must be made regarding the approach used to determine the interactions between all the electrons and nuclei. The choice of approach involves selecting an appropriate set of mathematical functions to represent the molecular orbitals, called a basis set[9,10] (Chapter 2, Section 2.1.4), and a theoretical method to solve the Schrödinger equation (Chapter 2, Section 2.1.1) or the Kohn–Sham equations (Chapter 2, Section 2.1.6). The choice of method and basis set are collectively referred to as the "level of theory". The level of theory can range from low to high based on the desired accuracy of the results produced. Unfortunately, the computational cost associated with performing QM simulations (CPU, memory, storage) and the time to complete them scale with the size of the chemical system (the number of atoms and electrons) and the level of theory.[11,12] A selection of commonly employed quantum mechanical methods is listed in Table 1 according to their accessibility on modern computers and accuracy in predicting an experimental property like bond dissociation enthalpy (BDE). It should be noted that this overview is by no means exhaustive and only serves the purpose of a qualitative illustration.

Table 1. List of different methods employed in quantum chemistry ranked by their typical accessible system size and accuracy.

Method	**Number of atoms (N)**	**Computational scaling**	**Mean absolute error in BDE***
full configuration interaction	$N \leq 2$	N!	< 0.1 kcal/mol
coupled-cluster theory	$N < 10$	$\geq N^6$	< 0.5 kcal/mol
Møller−Plesset perturbation theory and double-hybrid density functionals	$N < 100$	$\geq N^5$	< 2.5 kcal/mol
hybrid density functionals	$N < 1000$	N^4	< 5 kcal/mol
generalized-gradient approximation density functionals	$N < 10000$	N^3	< 15 kcal/mol
Hartree−Fock	$N < 10000$	N^3–N^4	< 30 kcal/mol

* See references 95−97 for more details on benchmarking against reference data, obtained experimentally or with theoretical procedures that can reproduce experimental results with high fidelity.

The solution of the Schrödinger equation produces a wavefunction (Chapter 2, Section 2.1.2) that provides a mathematical representation of the electronic structure of a molecule and can be used to predict any of its observable properties. The choice of the basis set determines the quality of the calculated wavefunction. In principle, the exact solution of the Schrödinger equation requires an infinite number of functions in the basis set (i.e., a complete basis set). Of course, in practice, a finite number of functions must be used. The smallest possible basis set, called a minimal basis set, contains one basis function per occupied atomic orbital for each atom in the chemical system and tends to produce the least accurate results. Basis sets can have 2, 3, or more functions per occupied orbital, with convergence to the complete basis set limit being achieved with around 5 functions per occupied orbital. The quality of the simulation improves with increasing basis set size, but the calculation times scale with the third power of the size of the basis set. Less-than-complete basis sets introduce errors in calculated properties, referred to as the basis set incompleteness error.[13,14]

The choice of method used to solve the Schrödinger equation also determines the accuracy of the results and computational cost. Assuming N is the size of the system under investigation, the approximate computational scaling of the conventional theoretical methods varies from N^3 to N^7 (or even higher). Methods based on configuration interaction and coupled-cluster theory (Chapter 2, Section 2.1.5) scale as N^6-N^7 and generally produce results in excellent agreement with experiment when used in conjunction with nearly complete basis sets.[15] They can produce chemical predictions that are accurate to a fraction of a kcal/mol.[16–18] In contrast, Møller–Plesset perturbation theory methods (Chapter 2, Section 2.1.5) scale as N^5 or greater and are relatively less accurate than configuration interaction and coupled-cluster theory methods. On the other hand, density-functional theory (DFT) methods, where the essential quantity is the electron density, are an alternative approach to wavefunction based methods (Chapter 2, Section 2.1.6). They can yield results within a few kcal/mol of the most accurate wavefunction based methods. In this context, double-hybrid density-functional theory methods scale as N^5 while other conventional DFT methods scale between N^3-N^4. The fastest all-electron wavefunction based method, i.e., the Hartree–Fock (HF) approach (Chapter 2, Section 2.1.3), also scales between N^3-N^4. Note that efficient HF algorithms make the method faster than most conventional DFT methods.

While configuration interaction and coupled-cluster theory methods with large basis sets yield more accurate results than other methods,[19] they are only applicable to chemical systems containing fewer than ten atoms.[20–22] The HF based methods with large basis sets can be used to model systems with hundreds of atoms but produce results in very poor agreement with the experiment.[15] The shortcomings of HF result from the inability of the method to describe the correlated motion of electrons. Alternatively, DFT based methods with large basis sets can also be used to model systems with hundreds of atoms and perform reasonably well but predict chemical properties with less accuracy than high-level theory. The shortcomings of DFT result mainly from the approximations in the underlying exchange-correlation functional (Chapter 2, Section 2.1.6).[23]

1.2 Research Motivation

The need for a level of theory that can be applied to the simulation of large chemical systems has driven the development of quantum mechanical methods where the aim is to find the right balance between computational efficiency and accuracy[24–29], and this is an active area of research[30–36]. QM based methods that balance computational cost and prediction accuracy can be useful in obtaining reliable structures, sampling conformational space, increasing exploration throughput, and predicting various molecular properties. The development of fast and accurate QM based methods that allow modeling systems with up to a few thousand atoms will have a benefit on various research areas by: (i) assisting in *a priori* design of supramolecular[37] architectures, (ii) enabling faster elucidation of proposed reaction pathways of enzyme-catalyzed biochemical reactions[38–41], (iii) allowing for quantum refinement of protein structures[42,43], (iv) extending the accessible system size and time scales in *ab initio* molecular dynamics[44,45] simulations of proteins[46–49], and (v) reducing the search space for faster identification of promising candidates in computer-aided materials[50–52] and drug[53–55] discovery/design.

1.3 Research Approach

A computationally inexpensive approach to model large chemical systems of interest containing up to a few thousand atoms is using HF and DFT based methods with a minimal basis set or basis sets containing two functions per occupied orbital (a double-ζ basis set). However, the low computational cost of such approaches comes at the expense of significant sacrifice in the prediction accuracy, thereby making them one of the most inaccurate techniques in quantum chemistry. Therefore, to predict chemical properties more accurately with minimal or double-ζ basis set HF and DFT methods, the underlying shortcomings[56] caused by the errors due to the incompleteness of basis set and the approximations in the HF and DFT methods must be mitigated with minimal computational overhead.

Over the last two decades, DiLabio and co-workers have shown that atom-centered potentials[57] (ACPs) offer a convenient means to improve the accuracy of HF and DFT methods. ACPs are equivalent to effective-core potentials[58–61] (Chapter 2, Section 2.1.4), which are used to describe the behavior of core electrons in heavy elements and are widely implemented in computational chemistry packages. However, unlike effective-core potentials, ACPs do not replace any core electrons. Instead, ACPs can be formulated to generate corrections to the energy and wavefunction calculated at a given, usually low-level of theory. An ACP based approach has some key advantages: (i) ACPs can be used in software that supports effective-core potentials without any code modification, (ii) ACPs can simultaneously correct for the multiple shortcomings inherent in many quantum mechanical approaches, (iii) it is also possible to use ACPs in conjunction with semi-empirical correction schemes that can also correct for deficiencies in low-level theory approaches, and (iv) application of ACPs leads to an increase in the computational cost of the underlying methods by only 10–30%.

The success of early ACP work in biomolecular[62–64] and silicon[65–67] modeling led to the development of a new generation of ACP based methods in the late 2000s and early 2010s.[68–72] By fitting to data obtained from high-level of theory, these newer ACPs were developed to correct the missing dispersion interactions in conventional DFT methods. These ACPs were applied to study several problems, including binding of various organic hydrocarbons on silicon surfaces[73–76], CO_2 adsorption on carbon nanotubes[77], binding in a series of thiophene and benzothiophene dimers[78], and the character of electronic transport in oligothiophene dimers[79]. One of the more interesting works carried out using ACP based methods was determining the lowest energy structure of asphaltene-type molecules and their related electronic properties, which predicted them to be a useful component for organic electronic devices, leading to a patent publication.[80,81]

Some system-specific ACP based methods were also developed in the DiLabio group to explore methane clusters[82] and water clusters[83]. In the past few years, development work has been carried out to design ACP based methods for a few elements to mitigate errors in predicting non-covalent interaction strengths and provide improved descriptions of covalent bonding and thermochemistry.[84,85] These works have served as a useful tool to develop a detailed understanding of the mechanism of some chemical reactions, in conjunction with various experimental studies. For example, they have been used to explore the influence of non-covalent interactions on the reaction kinetics and energetics of various chemical reactions where radical species are involved in hydrogen atom abstractions.[86–94]

More recently, Otero-de-la-Roza and DiLabio proposed their new ACP based approaches to rectify the severe basis set incompleteness error associated with small basis set DFT methods.[13,14] They demonstrated that ACPs fitted specifically to molecular properties obtained with complete basis set DFT could reproduce the properties when used with the same functional and a small basis set. In a different work[83], it was also shown that ACPs fitted to water cluster interaction energies indirectly brought the molecular dipole moment of water to an agreement with the experimental value to four significant digits, without specifically fitting to the dipole moment. This indicates that ACPs can successfully correct the molecular electronic distribution.

The success of these earlier works laid out a solid foundation for the work conducted as part of this dissertation, where new ACPs were developed to mitigate the shortcomings of minimal or double-ζ basis set HF and DFT methods. The work presented in this dissertation showcases a novel way of developing ACPs to overcome some major limitations associated with the previous generation ACPs. For example, earlier ACP development works focused mainly on improving the performance of DFT methods with moderate-sized or large basis sets, thereby limiting their application to systems with only a few hundred atoms. Such ACPs targeted a limited number of elements and molecular properties (mainly non-covalent) due to either scarcity of reference data in the literature or other complications. One major hurdle in the earlier development process was also the fitting procedure used to obtain the parameters of ACPs, which either required intensive manual labor or suffered from issues due to combinatorial explosion in selecting the best parameters. The ACPs developed herein overcome the limitations mentioned above. It is shown that the parameters of ACPs can be obtained by fitting to as many reference data points as wanted by using a more sophisticated and efficient fitting procedure. In fact, the fitting procedure utilized herein is shown to work with hundreds of thousands of data points and can extend the applicability of ACPs to more molecular properties than before for various applications.

1.4 Research Overview

This dissertation describes the development of fast and accurate QM based methods for modeling molecules with up to a few thousand atoms, including an accurate treatment of non-covalent interactions, chemical reaction kinetics, and thermochemistry. A suite of next-generation ACPs is presented that significantly mitigates the deficiencies of HF and DFT methods when used with minimal or double-ζ basis sets. It is anticipated that computational chemistry practitioners will be able to adopt the easy-to-use, fast, and accurate ACP based approaches developed as part of this dissertation in a wide range of (bio)chemistry problems.

The individual projects undertaken to achieve the goal of this dissertation are presented in Parts II–V, Chapters 3–9. Each of these parts address the various research objectives undertaken in this dissertation. The first objective is undertaking a proof-of-concept study of the ACP correction approach and is presented in Part II. Part III focusses on the second research objective of construction of new reference data sets to assist in further ACP development. The third objective addressed in Part IV is to conduct a comparative study of the ACP correction approach side-by-side to two semi-empirical correction techniques from literature. Finally, the objective achieved in Part V is the development of final ACPs for small basis set HF and DFT methods.

Part II consists of Chapter 3, where the capability of ACPs to accurately model structures and non-covalent interactions of large molecular systems is explored, particularly when used in conjunction with a minimal basis set HF method. In this chapter, the central hypothesis, whether ACPs can effectively reduce the errors in calculated molecular properties due to the incompleteness of the minimal basis set and the deficiencies of the HF method, is investigated. Extending the work in Chapter 3 to create more generally applicable ACPs is the purpose of Chapters 4–9. However, before this could be achieved, new accurate theoretical reference data was required, which could not be found in the literature. Consequently, four new data sets of molecular properties were constructed in Part III, Chapters 4–6. The molecular properties for which data sets were generated include peptide conformational energies (PEPCONF, Chapter 4), bond separation energies (BSE49, Chapter 5), reaction barrier heights (BH9, Chapter 6), and reaction energies (BH9-RE, Chapter 6). These new data sets, once generated, were then utilized for ACP development in Chapters 7–9.

In Part IV (Chapter 7), a detailed investigation was conducted to explore the performance of ACPs for modeling non-covalent interactions in the context of small basis set HF. In the same work, the performance of the ACPs was compared to some other semi-empirical correction schemes, and the various strengths and weaknesses associated with each strategy were identified. This work was important to understand if there was any value in developing ACPs for pairing with these correction schemes. Finally, Part V (Chapters 8 and 9) presents the development of the latest set of ACPs for ten elements (H, B, C, N, O, F, Si, P, S, Cl) designed for use with seven minimal or double-ζ HF and DFT methods. The work done in Chapters 8 and 9 is greatly motivated by the success of the ACP correction scheme explored in Chapters 3 and 7. ACPs developed for HF based approaches are presented in Chapter 8 and are designed to model mainly structures and non-covalent properties. ACPs developed for three common DFT based methods are presented in Chapter 9 and are designed to model both intermolecular interactions and thermochemistry. As a whole, the work conducted in this dissertation demonstrates the effectiveness of ACPs in improving the accuracy of low-level QM based methods so they can reproduce the accuracy of higher levels of theory to within a few kcal/mol. In doing so, they allow the efficient exploration of various applications involving large chemical and biological systems of interest with up to a few thousand atoms.

References

(1) Dykstra, C. E. *Theory and Applications of Computational Chemistry : The First Forty Years*; Elsevier, 2005.

(2) Krylov, A.; Windus, T. L.; Barnes, T.; Marin-Rimoldi, E.; Nash, J. A.; Pritchard, B.; Smith, D. G. A.; Altarawy, D.; Saxe, P.; Clementi, C.; et al. Perspective: Computational Chemistry Software and Its Advancement as Illustrated through Three Grand Challenge Cases for Molecular Science. *J. Chem. Phys.* **2018**, *149* (18), 180901.

(3) Hargittai, I. Models—Experiment—Computation: A History of Ideas in Structural Chemistry. In *Practical Aspects of Computational Chemistry I*; Springer Netherlands: Dordrecht, 2011; pp 1–31.

(4) Bytautas, L.; Matsunaga, N.; Nagata, T.; Gordon, M. S.; Ruedenberg, K. Accurate Ab Initio Potential Energy Curve of F2. III. The Vibration Rotation Spectrum. *J. Chem. Phys.* **2007**, *127* (20), 204313.

(5) The Nobel Prize in Chemistry 1998. NobelPrize.org. Nobel Media AB 2019 https://www.nobelprize.org/prizes/chemistry/1998/summary/ (accessed Oct 28, 2021).

(6) The Nobel Prize in Chemistry 2013. NobelPrize.org. Nobel Media AB 2019 https://www.nobelprize.org/prizes/chemistry/2013/summary/ (accessed Oct 28, 2021).

(7) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* **2016**, *116* (14), 7898–7936.

(8) Poltev, V. Molecular Mechanics: Principles, History, and Current Status. In *Handbook of Computational Chemistry*; Springer International Publishing: Cham, 2017; pp 21–67.

(9) Jensen, F. Atomic Orbital Basis Sets. *Wiley Interdiscip. Rev.Comput. Mol. Sci.* **2013**, *3* (3), 273–295. https://doi.org/10.1002/wcms.1123.

(10) Nagy, B.; Jensen, F. Basis Sets in Quantum Chemistry. In *Reviews in Computational Chemistry*; John Wiley & Sons, Ltd, 2017; pp 93–149.

(11) Ratcliff, L. E.; Mohr, S.; Huhs, G.; Deutsch, T.; Masella, M.; Genovese, L. Challenges in Large Scale Quantum Mechanical Calculations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *7* (1), e1290.

(12) Morokuma, K. New Challenges in Quantum Chemistry: Quests for Accurate Calculations for Large Molecular Systems. *Philos. Trans. R. Soc. London. Ser. A Math. Phys. Eng. Sci.* **2002**, *360* (1795), 1149–1164.

(13) Otero-De-La-Roza, A.; Dilabio, G. A. Improved Basis-Set Incompleteness Potentials for Accurate Density-Functional Theory Calculations in Large Systems. *J. Chem. Theory Comput.* **2020**, *16* (7), 4176–4191.

(14) Otero-de-la-Roza, A.; DiLabio, G. A. Transferable Atom-Centered Potentials for the Correction of Basis Set Incompleteness Errors in Density-Functional Theory. *J. Chem. Theory Comput.* **2017**, *13* (8), 3505–3524.

(15) Helgaker, T.; Jørgensen, P.; Olsen, J. Calibration of the Electronic-Structure Models. In *Molecular Electronic-Structure Theory*; John Wiley & Sons, Ltd: Chichester, UK, 2014; pp 817–883.

(16) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. HEAT: High Accuracy Extrapolated Ab Initio Thermochemistry. *J. Chem. Phys.* **2004**, *121* (23), 11599.

(17) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. W4 Theory for Computational Thermochemistry: In Pursuit of Confident Sub-KJ/Mol Predictions. *J. Chem. Phys.* **2006**, *125* (14), 144108.

(18) Sylvetsky, N.; Peterson, K. A.; Karton, A.; Martin, J. M. L. Toward a W4-F12 Approach: Can Explicitly Correlated and Orbital-Based Ab Initio CCSD(T) Limits Be Reconciled? *J. Chem. Phys.* **2016**, *144* (21), 214101.

(19) Ke dziera, D.; Kaczmarek-Kedziera, A. Remarks on Wave Function Theory and Methods. In *Handbook of Computational Chemistry*; Springer International Publishing: Cham, 2017; pp 123–171.

(20) Hohenstein, E. G.; Sherrill, C. D. Wavefunction Methods for Noncovalent Interactions. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2* (2), 304–326.

(21) Černý, J.; Hobza, P. Non-Covalent Interactions in Biomacromolecules. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5291–5303.

(22) Riley, K. E.; Hobza, P. Noncovalent Interactions in Biochemistry. *Wiley Interdiscip. Rev. Comput Mol Sci* **2011**, *1*, 3–17.

(23) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112* (1), 289–320.

(24) Warshel, A. Multiscale Modeling of Biological Functions: From Enzymes to Molecular Machines (Nobel Lecture). *Angew. Chemie Int. Ed.* **2014**, *53* (38), 10020–10031.

(25) Cui, Q. Perspective: Quantum Mechanical Methods in Biochemistry and Biophysics. *J. Chem. Phys.* **2016**, *145* (14), 140901.

(26) Merz, K. M. Using Quantum Mechanical Approaches to Study Biological Systems. *Acc. Chem. Res.* **2014**, *47* (9), 2804–2811.

(27) Cole, D. J.; Hine, N. D. M. Applications of Large-Scale Density Functional Theory in Biology. *J. Phys. Condens. Matter* **2016**, *28* (39), 393001.

(28) Gordon, M. S.; Mullin, J. M.; Pruitt, S. R.; Roskop, L. B.; Slipchenko, L. V.; Boatz, J. A. Accurate Methods for Large Molecular Systems †. *J. Phys. Chem. B* **2009**, *113* (29), 9646–9663.

(29) Li, X.; Chung, L. W.; Morokuma, K. Modeling Photobiology Using Quantum Mechanics and Quantum Mechanics/Molecular Mechanics Calculations. In *Computational Methods for Large Systems*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2011; pp 397–433.

(30) Hofer, T. S. From Macromolecules to Electrons—Grand Challenges in Theoretical and Computational Chemistry. *Front. Chem.* **2013**, *1*, 6.

(31) Grimme, S.; Schreiner, P. R. Computational Chemistry: The Fate of Current Methods and Future Challenges. *Angew. Chemie Int. Ed.* **2018**, *57* (16), 4170–4176.

(32) Houk, K. N.; Liu, F. Holy Grails for Computational Organic Chemistry and Biochemistry. *Acc. Chem. Res.* **2017**, *50* (3), 539–543.

(33) Neese, F.; Atanasov, M.; Bistoni, G.; Maganas, D.; Ye, S. Chemistry and Quantum Mechanics in 2019: Give Us Insight and Numbers. *J. Am. Chem. Soc.* **2019**, *141* (7), 2814–2824.

(34) Brunk, E.; Ashari, N.; Athri, P.; Campomanes, P.; de Carvalho, F. F.; Curchod, B. F. E.; Diamantis, P.; Doemer, M.; Garrec, J.; Laktionov, A.; et al. Pushing the Frontiers of First-Principles Based Computer Simulations of Chemical and Biological Systems. *Chim. Int. J. Chem.* **2011**, *65* (9), 667–671.

(35) Sherrill, C. D. Frontiers in Electronic Structure Theory. *J. Chem. Phys.* **2010**, *132* (11), 110902.

(36) Bachrach, S. M. Challenges in Computational Organic Chemistry. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4* (5), 482–487.

(37) Lehn, J. *Supramolecular Chemistry: Concepts and Perspectives*; Wiley VCH, 1995.

(38) Schramm, V. L. Enzymatic Transition States, Transition-State Analogs, Dynamics, Thermodynamics, and Lifetimes. *Ann. Rev. Biochem.* **2011**, *80*, 703–732.

(39) Vaissier Welborn, V.; Head-Gordon, T. Computational Design of Synthetic Enzymes. *Chem. Rev.* **2018**, acs.chemrev.8b00399.

(40) Marcos, E.; Silva, D.-A. Essentials of *de Novo* Protein Design: Methods and Applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8* (6), e1374.

(41) Zanghellini, A. De Novo Computational Enzyme Design. *Curr. Opin. Biotechnol.* **2014**, *29*, 132–138.

(42) Goerigk, L.; Collyer, C. A.; Reimers, J. R. Recommending Hartree–Fock Theory with London-Dispersion and Basis-Set-Superposition Corrections for the Optimization or Quantum Refinement of Protein Structures. *J. Phys. Chem. B* **2014**, *118* (50), 14612–14626.

(43) Hsiao, Y. W.; Sanchez-Garcia, E.; Doerr, M.; Thiel, W. Quantum Refinement of Protein Structures: Implementation and Application to the Red Fluorescent Protein DsRed.M1. *J. Phys. Chem. B* **2010**, *114* (46), 15413–15423.

(44) Paquet, E.; Viktor, H. L. Computational Methods for Ab Initio Molecular Dynamics. *Adv. Chem.* **2018**, *2018*, 1–14.

(45) Marx, D.; Hutter, J. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*; Cambridge University Press, 2009.

(46) Dal Peraro, M.; Raugei, S.; Carloni, P.; Klein, M. L. Solute-Solvent Charge Transfer in Aqueous Solution. *ChemPhysChem* **2005**, *6* (9), 1715–1718.

(47) Ufimtsev, I. S.; Luehr, N.; Martinez, T. J. Charge Transfer and Polarization in Solvated Proteins from Ab Initio Molecular Dynamics. *J. Phys. Chem. Lett.* **2011**, *2* (14), 1789–1793.

(48) Ufimtsev, I. S.; Martinez, T. J. Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. *J. Chem. Theory Comput.* **2009**, *5* (10), 2619–2628.

(49) Wei, D.; Guo, H.; Salahub, D. R. Conformational Dynamics of an Alanine Dipeptide Analog: An *Ab Initio* Molecular Dynamics Study. *Phys. Rev. E* **2001**, *64* (1), 011907.

(50) Lu, Z. Computational Discovery of Energy Materials in the Era of Big Data and Machine Learning: A Critical Review. *Mater. Reports Energy* **2021**, *1* (3), 100047.

(51) Oganov, A. R.; Saleh, G.; Kvashnin A. G. *Computational Materials Discovery*; Royal Society of Chemistry, 2018.

(52) Oganov, A. R.; Pickard, C. J.; Zhu, Q.; Needs, R. J. Structure Prediction Drives Materials Discovery. *Nat. Rev. Mater. 2019 45* **2019**, *4* (5), 331–348.

(53) Mulholland, A. J.; Amaro, R. E. COVID19 - Computational Chemists Meet the Moment. *J. Chem. Inf. Model.* **2020**, *60* (12), 5724–5726.

(54) Gurung, A. B.; Ali, M. A.; Lee, J.; Farah, M. A.; Al-Anazi, K. M. An Updated Review of Computer-Aided Drug Design and Its Application to COVID-19. *Biomed Res. Int.* **2021**, *2021*, 8853056.

(55) Medina-Franco, J. L. Grand Challenges of Computer-Aided Drug Design: The Road Ahead. *Front. Drug Discov.* **2021**, *0*, 2.

(56) Sure, R.; Brandenburg, J. G.; Grimme, S. Small Atomic Orbital Basis Set First-Principles Quantum Chemical Methods for Large Molecular and Periodic Systems: A Critical Analysis of Error Sources. *ChemistryOpen* **2016**, *5* (2), 94–109.

(57) DiLabio, G. A. Atom-Centered Potentials for Noncovalent Interactions and Other Applications. In *Non-Covalent Interactions in Quantum Chemistry and Physics: Theory and Applications*; Elsevier Inc., 2017; pp 221–240.

(58) Cao, X.; Dolg, M. Pseudopotentials and Modelpotentials. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (2), 200–210.

(59) Dolg, M.; Cao, X. Relativistic Pseudopotentials: Their Development and Scope of Applications. *Chem. Rev.* **2012**, *112* (1), 403–480.

(60) Dolg, M. Relativistic Effective Core Potentials. In *Handbook of Relativistic Quantum Chemistry*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2017; pp 449–478.

(61) Cundari, T. R.; Benson, M. T.; Lutz, M. L.; Sommerer, S. O. Effective Core Potential Approaches to the Chemistry of the Heavier Elements; John Wiley & Sons, Ltd, 2007; pp 145–202.

(62) Johnson, E. R.; DiLabio, G. A. Convergence of Calculated Nuclear Magnetic Resonance Chemical Shifts in a Protein with Respect to Quantum Mechanical Model Size. *J. Mol. Struct. THEOCHEM* **2009**, *898* (1–3), 56–61.

(63) Moon, S.; Christiansen, P. A.; DiLabio, G. A. Quantum Capping Potentials with Point Charges: A Simple QM/MM Approach for the Calculation of Large-Molecule NMR Shielding Tensors. *J. Chem. Phys.* **2004**, *120* (19), 9080–9086.

(64) DiLabio, G. A.; Hurley, M. M.; Christiansen, P. A. Simple One-Electron Quantum Capping Potentials for Use in Hybrid QM/MM Studies of Biological Molecules. *J. Chem. Phys.* **2002**, *116* (22), 9578–9584.

(65) Dogel, I. A.; Dogel, S. A.; Pitters, J. L.; DiLabio, G. A.; Wolkow, R. A. Chemical Methods for the Hydrogen Termination of Silicon Dangling Bonds. *Chem. Phys. Lett.* **2007**, *448* (4–6), 237–242.

(66) DiLabio, G. A.; Dogel, S. A.; Wolkow, R. A. A Simple and Accurate Approach for Calculating the Vibration Spectra of Molecules on Surfaces: Comparisons to High Resolution Electron Energy Loss Data for Ethylene on Silicon. *Surf. Sci.* **2006**, *600* (16), L209–L213.

(67) DiLabio, G. A.; Wolkow, R. A.; Johnson, E. R. Efficient Silicon Surface and Cluster Modeling Using Quantum Capping Potentials. *J. Chem. Phys.* **2005**, *122* (4), 044708.

(68) Torres, E.; DiLabio, G. A. A (Nearly) Universally Applicable Method for Modeling Noncovalent Interactions Using B3LYP. *J. Phys. Chem. Lett.* **2012**, *3* (13), 1738–1744.

(69) Mackie, I. D.; DiLabio, G. A. Accurate Dispersion Interactions from Standard Density-Functional Theory Methods with Small Basis Sets. *Phys. Chem. Chem. Phys.* **2010**, *12* (23), 6092.

(70) Mackie, I. D.; DiLabio, G. A. Interactions in Large, Polyaromatic Hydrocarbon Dimers: Application of Density Functional Theory with Dispersion Corrections. *J. Phys. Chem. A* **2008**, *112* (43), 10968–10976.

(71) DiLabio, G. A. Accurate Treatment of van Der Waals Interactions Using Standard Density Functional Theory Methods

with Effective Core-Type Potentials: Application to Carbon-Containing Dimers. *Chem. Phys. Lett.* **2008**, *455* (4–6), 348–353.

(72) DiLabio, G. A.; Koleini, M.; Torres, E. Extension of the B3LYP–Dispersion-Correcting Potential Approach to the Accurate Treatment of Both Inter- and Intra-Molecular Interactions. *Theor. Chem. Acc.* **2013**, *132* (10), 1389.

(73) DiLabio, G. A.; Johnson, E. R.; Pitters, J. Pentacene Binds Strongly to Hydrogen-Terminated Silicon Surfaces Via Dispersion Interactions. *J. Phys. Chem. C* **2009**, *113* (23), 9969–9973.

(74) Zikovsky, J.; Dogel, S. A.; Salomons, M. H.; Pitters, J. L.; DiLabio, G. A.; Wolkow, R. A. Indications of Field-Directing and Self-Templating Effects on the Formation of Organic Lines on Silicon. *J. Chem. Phys.* **2011**, *134* (11), 114707.

(75) Sinha, S.; DiLabio, G. A.; Wolkow, R. A. Experimental and Theoretical Exploration of the Anisotropy of Styrene Diffusion on Hydrogen Terminated Si(100)-2 × 1. *J. Phys. Chem. C* **2010**, *114* (16), 7364–7371.

(76) Johnson, E. R.; DiLabio, G. A. Theoretical Study of Dispersion Binding of Hydrocarbon Molecules to Hydrogen-Terminated Silicon(100)-2×1. *J. Phys. Chem. C* **2009**, *113* (14), 5681–5689.

(77) Mackie, I. D.; DiLabio, G. A. CO $_2$ Adsorption by Nitrogen-Doped Carbon Nanotubes Predicted by Density-Functional Theory with Dispersion-Correcting Potentials. *Phys. Chem. Chem. Phys.* **2011**, *13* (7), 2780–2787.

(78) Mackie, I. D.; McClure, S. A.; DiLabio, G. A. Binding in Thiophene and Benzothiophene Dimers Investigated By Density Functional Theory with Dispersion-Correcting Potentials. *J. Phys. Chem. A* **2009**, *113* (18), 5476–5484.

(79) McClure, S. A.; Buriak, J. M.; DiLabio, G. A. Transport Properties of Thiophenes: Insights from Density-Functional Theory Modeling Using Dispersion-Correcting Potentials. *J. Phys. Chem. C* **2010**, *114* (24), 10952–10961.

(80) Mackie, I. D.; DiLabio, G. A. Importance of the Inclusion of Dispersion in the Modeling of Asphaltene Dimers. *Energy & Fuels* **2010**, *24* (12), 6468–6475.

(81) DiLabio, Gino A.; Mackie, Iain; Dettman, Heather D. Asphaltene Components as Organic Electronic Materials, US Patent, 9,065,059, 2015.

(82) Torres, E.; DiLabio, G. A. Density-Functional Theory with Dispersion-Correcting Potentials for Methane: Bridging the Efficiency and Accuracy Gap between High-Level Wave Function and Classical Molecular Mechanics Methods. *J. Chem. Theory Comput.* **2013**, *9* (8), 3342–3349.

(83) Holmes, J. D.; Otero-de-la-Roza, A.; DiLabio, G. A. Accurate Modeling of Water Clusters with Density-Functional Theory Using Atom-Centered Potentials. *J. Chem. Theory Comput.* **2017**, *13* (9), 4205–4215.

(84) van Santen, J. A.; DiLabio, G. A. Dispersion Corrections Improve the Accuracy of Both Noncovalent and Covalent Interactions Energies Predicted by a Density-Functional Theory Approximation. *J. Phys. Chem. A* **2015**, *119* (25), 6703–6713.

(85) DiLabio, G. A.; Koleini, M. Dispersion-Correcting Potentials Can Significantly Improve the Bond Dissociation Enthalpies and Noncovalent Binding Energies Predicted by Density-Functional Theory. *J. Chem. Phys.* **2014**, *140* (18), 18A542.

(86) Salamone, M.; Milan, M.; DiLabio, G. A.; Bietti, M. Reactions of the Cumyloxyl and Benzyloxyl Radicals with Tertiary Amides. Hydrogen Abstraction Selectivity and the Role of Specific Substrate-Radical Hydrogen Bonding. *J. Org. Chem.* **2013**, *78* (12), 5909–5917.

(87) Bietti, M.; Salamone, M.; DiLabio, G. A.; Jockusch, S.; Turro, N. J. Kinetic Solvent Effects on Hydrogen Abstraction from Phenol by the Cumyloxyl Radical. Toward an Understanding of the Role of Protic Solvents. *J. Org. Chem.* **2012**, *77* (3), 1267–1272.

(88) van Santen, J. A.; Rahemtulla, S.; Salamone, M.; Bietti, M.; DiLabio, G. A. A Computational and Experimental Re-Examination of the Reaction of the Benzyloxyl Radical with DMSO. *Comput. Theor. Chem.* **2016**, *1077*, 74–79.

(89) DiLabio, G. A.; Franchi, P.; Lanzalunga, O.; Lapi, A.; Lucarini, F.; Lucarini, M.; Mazzonna, M.; Prasad, V. K.; Ticconi, B. Hydrogen Atom Transfer (HAT) Processes Promoted by the Quinolinimide- *N* -Oxyl Radical. A Kinetic and Theoretical Study. *J. Org. Chem.* **2017**, *82* (12), 6133–6141.

(90) Mackie, I. D.; DiLabio, G. A. Ring-Opening Radical Clock Reactions: Many Density Functionals Have Difficulty Keeping Time. *Org. Biomol. Chem.* **2011**, *9* (9), 3158.

(91) D'Alfonso, C.; Bietti, M.; DiLabio, G. A.; Lanzalunga, O.; Salamone, M. Reactions of the Phthalimide *N* -Oxyl Radical

(PINO) with Activated Phenols: The Contribution of π-Stacking Interactions to Hydrogen Atom Transfer Rates. *J. Org. Chem.* **2013**, *78* (3), 1026–1037.

(92) Mazzonna, M.; Bietti, M.; DiLabio, G. A.; Lanzalunga, O.; Salamone, M. Importance of π-Stacking Interactions in the Hydrogen Atom Transfer Reactions from Activated Phenols to Short-Lived *N* -Oxyl Radicals. *J. Org. Chem.* **2014**, *79* (11), 5209–5218.

(93) Coyle, J. P.; Johnson, P. A.; DiLabio, G. A.; Barry, S. T.; Müller, J. Gas-Phase Thermolysis of a Guanidinate Precursor of Copper Studied by Matrix Isolation, Time-of-Flight Mass Spectrometry, and Computational Chemistry. *Inorg. Chem.* **2010**, *49* (6), 2844–2850.

(94) Salamone, M.; DiLabio, G. A.; Bietti, M. Reactions of the Cumyloxyl and Benzyloxyl Radicals with Strong Hydrogen Bond Acceptors. Large Enhancements in Hydrogen Abstraction Reactivity Determined by Substrate/Radical Hydrogen Bonding. *J. Org. Chem.* **2012**, *77* (23), 10479–10487.

(95) Feng, Y; Liu, L; Wang, J.-T.; Huang, H; Guo, Q.-X. Assessment of Experimental Bond Dissociation Energies Using Composite ab Initio Methods and Evaluation of the Performances of Density Functional Methods in the Calculation of Bond Dissociation Energies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2005–2013.

(96) Chan, B.; Radom, L. BDE261: A Comprehensive Set of High-Level Theoretical Bond Dissociation Enthalpies. *J. Phys. Chem. A.* **2012**, *116*, 4975–4986.

(97) Zhao, Y.; Ng, H. T.; Peverati, R.; Truhlar, D. G. Benchmark Database for Ylidic Bond Dissociation Energies and Its Use for Assessments of Electronic Structure Methods. *J. Chem. Theory Comput.* **2012**, *8*, 2824–2834.

Chapter 2

Theoretical Background

This chapter presents the details of various computational methods and techniques employed during the PhD studies. Readers can skip this chapter without sacrificing the continuity of the dissertation as the methodological details relevant to the various projects have been presented in each published/submitted research chapter.

The field of computational chemistry is based on the application of quantum mechanics, a discipline that predates the first computers by many decades, to the solution of chemical problems. The field has progressed significantly since the pioneering foundational work done in the early twentieth century that led to development of quantum mechanics. Quantum mechanics arose from an effort to understand various physical phenomena occurring at a microscopic scale that classical physics was unable to explain. These phenomena include the frequency dependence of the black-body radiation, the photoelectric effect, and the discrete nature of the emission spectrum of the hydrogen atom. The concepts associated with quantum mechanics can be found in any introductory or advanced physical chemistry textbook and are not discussed here. An interested reader is referred to references 1–4 for more information.

In the context of chemistry, quantum mechanics is used to describe the electronic structure of atoms and molecules. Quantum chemistry can be applied to predict molecular properties from the electronic distribution, such as bond strengths and chemical reaction energies. The application of quantum chemistry typically involves the solution of very challenging partial differential equations, so early applications in this field were restricted to atomic, diatomic, or symmetrical molecules. However, with the advent of computers and the increase in computing power in the last decades, quantum chemistry methods are now applicable using software packages.[5] The development of quantum chemistry methods and their application to problems in chemistry constitutes the field of computational chemistry. Modern-day computational chemistry can now be applied to study complex molecular systems of real and practical interest. The next sections provide an overview of the various theoretical methods and techniques rooted in quantum mechanics that form the foundations of this dissertation. For all the methods and techniques described in this chapter, a handful of excellent books and review articles are available. The interested reader is referred to references 6–10 and other suggested references in the sections below for more details.

2.1 Computational Chemistry Methods

2.1.1 The Molecular Hamiltonian

One of the key equations in quantum mechanics is the partial differential eigenvalue equation called the Schrödinger equation.[11,12] In quantum chemistry, the most fundamental problem is to find solutions to the non-relativistic time-independent Schrödinger equation:

$$\hat{H}|\Psi\rangle = E|\Psi\rangle \tag{2.1}$$

where $\hat{H}$ is a Hermitian operator called the Hamiltonian of a molecular system and $|\Psi\rangle$ (represented in standard Dirac bra-ket matrix notation[13]) is the wavefunction. The wavefunction characterizes the motion of electrons and nuclei for a given molecular state and all observable molecular properties may be derived from it. A wavefunction is an eigenvector of the Hamiltonian operator with the energy E as its eigenvalue. Since $\hat{H}$ is Hermitian, all eigenvalues are real, and the corresponding eigenvectors can be chosen as orthonormal to each other.

For a system of N electrons and M nuclei, the molecular Hamiltonian is expressed as a sum of five terms:

$$\hat{H} = \hat{T}_e + \hat{T}_n + \hat{V}_{en} + \hat{V}_{ee} + \hat{V}_{nn} \tag{2.2}$$

The terms in Equation 2.2 represent the kinetic energy of the electrons ($\hat{T}_e$) and nuclei ($\hat{T}_n$), and the potential energies arising from the electrostatic interactions between electrons and nuclei ($\hat{V}_{en}$), electrons and electrons ($\hat{V}_{ee}$), and nuclei and nuclei ($\hat{V}_{nn}$). The molecular Hamiltonian from Equation 2.2 is expressed in atomic units ($\frac{\mathrm{h}}{2\pi} = \mathrm{e} = \mathrm{m_e} = 4\pi\varepsilon_0 = 1\ \mathrm{a.u.}$) as:

$$\begin{aligned}\hat{H} = -\sum_{a=1}^{N}\frac{1}{2}\nabla_a^2 - \sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2 - \sum_{a=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{|\,\vec{r}_a - \vec{R}_A\,|} + \sum_{a=1}^{N-1}\sum_{b>a}^{N}\frac{1}{|\,\vec{r}_a - \vec{r}_b\,|} \\ + \sum_{A=1}^{M-1}\sum_{B>A}^{M}\frac{Z_A Z_B}{|\,\vec{R}_A - \vec{R}_B\,|}\end{aligned} \tag{2.3}$$

where a, b label electrons and A, B label nuclei, $\vec{r}_a$ are the electron coordinates, $\vec{R}_A$ are the nuclear coordinates, Z_A are the atomic numbers, and M_A are the atomic masses. The Laplacian operators ∇_a^2 and ∇_A^2 represent differentiation with respect to the coordinates of the a-th electron and the A-th nucleus.

Solving Equation 2.1 requires solving a $3(N + M)$-dimensional partial differential equation as both the Hamiltonian and the wavefunction depend on the N electronic and M nuclear coordinates. This is an intractable problem for all except the simplest systems. As a result, a simplification of the Schrödinger equation is required, which brings us to the Born−Oppenheimer approximation.[14] In this approximation, the comparatively slow motion of nuclei is decoupled from the motion of the fast-moving electrons. This can be justified by the fact that nuclei are heavier than electrons. Therefore, the motion of electrons is assumed to be "instantaneous" relative to the nuclei, resulting in the assumption that the electrons move in the electric field created by stationary nuclei. The Born−Oppenheimer approximation is nearly always applied in quantum chemical calculations. This approximation simplifies Equation 2.3 because decoupling the electronic and nuclear degrees of freedom leads to separate equations for the two types of particles. In the equation corresponding to the electrons, the nuclei are fixed and therefore the nuclear-nuclear repulsion term ($\hat{V}_{nn}$) from Equation 2.2 gives a constant contribution and the nuclear kinetic energy term disappears, i.e., $\hat{T}_n = 0$. This simplified Schrödinger equation is referred to as the electronic Schrödinger equation and is expressed as:

$$\hat{H}_e|\Psi_e\rangle = E_e|\Psi_e\rangle \tag{2.4}$$

where $\hat{H}_e$, $|\Psi_e\rangle$, and E_e are the electronic Hamiltonian, electronic wavefunction, and electronic energy, respectively. The electronic Hamiltonian adopts the following form:

$$\hat{H}_e = \hat{T}_e + \hat{V}_{en} + \hat{V}_{ee} \tag{2.5}$$

or,

$$\hat{H}_e = -\sum_{a=1}^{N}\frac{1}{2}\nabla_a^2 - \sum_{a=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{|\vec{r}_a - \vec{R}_A|} + \sum_{a=1}^{N-1}\sum_{b>a}^{N}\frac{1}{|\vec{r}_a - \vec{r}_b|} \tag{2.6}$$

The $\hat{H}_e$ operator depends on N electronic coordinates only, but there is one electronic Schrödinger equation for every arrangement of nuclei in space. Therefore, the Schrödinger equation that needs to be solved now is an eigenvalue problem in $3N$ dimensions only, compared to $3(N + M)$ before applying the Born−Oppenheimer approximation. The electronic energy plus the nuclear-nuclear repulsion term ($\hat{V}_{nn}$) as a function of the nuclear positions is known as the potential energy surface (see Section 2.2.1). The potential energy surface can be calculated by changing the nuclear coordinates, solving the electronic Schrödinger equation to obtain the electronic energy, and then adding the potential energy corresponding to $\hat{V}_{nn}$ to yield the total energy of the system. Unfortunately, even after Born−Oppenheimer approximation, Equation 2.4 can be only exactly solved for systems with very small number of electrons. Therefore, additional approximations are required to proceed to multi-electron systems as described next.

2.1.2 The Molecular Wavefunction

Before considering the nature of the wavefunction of a multi-electron system, we examine the wavefunction of a single electron system. The wavefunction of a single electron is referred to as an orbital. In molecules, the wavefunction of a single electron is referred to as a molecular orbital (MO). Because electrons are half-spin particles, and therefore have two possible spin states, the correct representation of the spatial distribution of an electron requires both a spatial and spin components for the molecular orbital. The spatial part is represented as $\psi(\vec{r})$ which is a function of the space coordinates $\vec{r} = x\hat{\imath} + y\hat{\jmath} + z\hat{k}$. The spin part is represented as $\alpha(\omega)$ for spin-up (↑) and $\beta(\omega)$ for spin-down (↓) which are functions of the spin coordinate ω. The combination of spatial and spin components is called a spin orbital and represented as $\chi(\vec{R})$, where $\vec{R} = \{\vec{r}, \omega\}$ indicates both space and spin coordinates. Because electrons are fermions, two electrons in the same system cannot occupy the same spin orbital. A single $\psi(\vec{r})$ can give two different spin orbitals when combined with the up or down spin functions, reflecting that the same spatial orbital can accommodate exactly one electron pair, each electron having opposite spin. In a multi-electron molecule, a set of N spatial orbitals $\{\psi_i(\vec{r})|\ i = 1, 2, 3, \ldots, N\}$ yields a set of *2N* spin orbitals $\{\chi_i(\vec{R})|\ i = 1, 2, 3, \ldots, N, \ldots, 2N\}$ as follows:

$$\begin{aligned} \chi_{2i-1}(\vec{R}) &= \psi_i(\vec{r})\alpha(\omega) \\ \chi_{2i}(\vec{R}) &= \psi_i(\vec{r})\beta(\omega) \end{aligned} \tag{2.7}$$

and these can be occupied by the system's electrons using the familiar aufbau principle.

The wavefunction for an N-electron system is a function of the coordinates of the N electrons, i.e., $\vec{R}_1, \vec{R}_2, \vec{R}_3, \ldots, \vec{R}_N$ and can be represented as $\Psi_e(\vec{R}_1, \vec{R}_2, \vec{R}_3, \ldots, \vec{R}_N)$. In the simplest approximation, the N-electron wavefunction is represented as a simple product of N spin orbitals**.** This approximation was first used by Hartree, and therefore the resulting wavefunction is referred to as the "Hartree product"[15]:

$$\Psi_e(\vec{R}_1, \vec{R}_2, \vec{R}_3, \ldots, \vec{R}_N) \approx \chi_1(\vec{R}_1)\chi_2(\vec{R}_2)\chi_3(\vec{R}_3) \cdots \chi_N(\vec{R}_N) \tag{2.8}$$

Because electrons in a molecule are indistinguishable fermions, quantum mechanics postulates that the N-electron wavefunction must be antisymmetric with respect to the exchange of any two electrons. This is known as the antisymmetry principle. It is mathematically expressed as follows:

$$\Psi_e(\vec{R}_1, \dots, \vec{R}_i, \dots, \vec{R}_j, \dots, \vec{R}_N) = -\Psi_e(\vec{R}_1, \dots, \vec{R}_j, \dots, \vec{R}_i, \dots, \vec{R}_N) \tag{2.9}$$

The N-electron wavefunction must satisfy the Schrödinger equation as well as obey the antisymmetry principle. The Hartree product is not antisymmetric with respect to the interchange of two electrons and thus is not a valid N-electron wavefunction. Nonetheless, the antisymmetry requirement can be easily realized with a set of N spin orbitals by writing the approximate N-electron wavefunction as a "Slater determinant" (this approximation was first utilized independently by Heisenberg[16] and Dirac[17] and is named after Slater[18]). For a N-electron system where the electrons occupy N spin orbitals $(\chi_1(\vec{R}_1), \chi_2(\vec{R}_2), \chi_3(\vec{R}_3), \cdots, \chi_N(\vec{R}_N))$, the Slater determinant is:

$$\left|\Psi_e(\vec{R}_1, \vec{R}_2, \vec{R}_3, \dots, \vec{R}_N)\right\rangle \approx \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(\vec{R}_1) & \chi_2(\vec{R}_1) & \cdots & \chi_N(\vec{R}_1) \\ \chi_1(\vec{R}_2) & \chi_2(\vec{R}_2) & \cdots & \chi_N(\vec{R}_2) \\ \chi_1(\vec{R}_3) & \chi_2(\vec{R}_3) & \cdots & \chi_N(\vec{R}_3) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(\vec{R}_N) & \chi_2(\vec{R}_N) & \cdots & \chi_N(\vec{R}_N) \end{vmatrix} \tag{2.10}$$

This $N \times N$ Slater determinant consists of N rows labeled by the electron coordinates and N columns labeled by the spin orbitals. The $(N!)^{-1/2}$ in Equation 2.10 is a normalization factor. The Slater determinant is a linear combination of several Hartree products that obeys the antisymmetric principle by construction. Equation 2.10 can be re-written as a sum of $N!$ Hartree products by applying the algebraic definition of a determinant:

$$\Psi_e(\vec{R}_1, \vec{R}_2, \vec{R}_3, \dots, \vec{R}_N) \approx \frac{1}{\sqrt{N!}} \sum_{j}^{N!} (-1)^{p_j} \hat{P}_j \prod_{i}^{n} \chi_i(\vec{R}_i) \tag{2.11}$$

where $N!$ is the number of unique permutations possible for the N factors in the Hartree product (Equation 2.8). The $\hat{P}_j$ in Equation 2.11 is the permutation operator. Its effect is to apply permutation j on the electron labels $\vec{R}_1, \vec{R}_2, \vec{R}_3 \dots, \vec{R}_N$. The $(-1)^{p_j}$ prefactor is the parity of the permutation, where p_j is the number of transpositions contained in permutation j.

The Slater determinant satisfies the antisymmetric principle by construction because interchanging the coordinates of two electrons corresponds to interchanging two rows of the determinant, thereby introducing a negative sign. Due to the determinantal form of the approximated wavefunction, the Pauli exclusion principle[19] is also obeyed and exchange effects are incorporated because having two electrons occupy the same spin orbital corresponds to having two equal columns in the determinant, thereby making the determinant zero. For simplicity, the Slater determinant from Equation 2.10 or 2.11 will be represented just by its diagonal elements and the $(N!)^{-1/2}$ normalization constant will be dropped:

$$\Psi_e\left(\vec{R}_1, \vec{R}_2, \vec{R}_3, \dots, \vec{R}_N\right) \approx \left|\chi_1\left(\vec{R}_1\right)\chi_2\left(\vec{R}_2\right)\chi_3\left(\vec{R}_3\right) \cdots \chi_N\left(\vec{R}_N\right)\right\rangle \tag{2.12}$$

Equation 2.12 can be further shortened by assuming that we always choose the electron coordinates to be in the order $\vec{R}_1, \vec{R}_2, \vec{R}_3, \dots, \vec{R}_N$:

$$\Psi_e\left(\vec{R}_1, \vec{R}_2, \vec{R}_3, \dots, \vec{R}_N\right) \approx \left|\chi_1\chi_2\chi_3 \cdots \chi_N\right\rangle \tag{2.13}$$

2.1.3 The Hartree–Fock Method

The ground-state electronic energy (E_e) associated with the exact electronic wavefunction $|\Psi_e\rangle$ can be obtained from the expectation value of the electronic Hamiltonian operator $\widehat{H}_e$ (see Equation 2.5 and 2.6) when it acts on the wavefunction Ψ_e, which is expressed in the Dirac bra-ket notation[13] as:

$$E_e = \left\langle \Psi_e \middle| \widehat{H}_e \middle| \Psi_e \right\rangle \tag{2.14}$$

The variational principle[3] states that the expectation value $(\tilde{E}_e)$ of any trial electronic wavefunction $(\widetilde{\Psi}_e)$ is always higher than the exact electronic energy (E_e) associated with the true electronic wavefunction (Ψ_e):

$$\tilde{E}_e = \left\langle \widetilde{\Psi}_e \middle| \widehat{H}_e \middle| \widetilde{\Psi}_e \right\rangle \geq \left\langle \Psi_e \middle| \widehat{H}_e \middle| \Psi_e \right\rangle = E_e \tag{2.15}$$

In addition, $\tilde{E}_e$ is equal to E_e if and only if $\widetilde{\Psi}_e$ is equal to Ψ_e. A common procedure employed to find approximate solutions to the electronic Schrödinger equation (see Equation 2.4) begins with a trial wavefunction $\widetilde{\Psi}_e$ and then uses the variational principle to minimize the energy expectation value $\left\langle \widetilde{\Psi}_e \middle| \widehat{H}_e \middle| \widetilde{\Psi}_e \right\rangle$ as a function of some parameters in $\widetilde{\Psi}_e$. In the simplest approximation, the chosen trial wavefunction is a Slater determinant (represented as Φ_0), and the energy is minimized with respect to the choice of trial MOs $\left(\chi_i\left(\vec{R}_i\right)\right)$. If one performs the variational minimization iteratively until self-consistency is achieved one obtains the best MOs according to the variational principle, the optimal determinantal wavefunction $\widetilde{\Psi}_e$ and its associated electronic energy $\tilde{E}_e$, which is necessarily higher than the exact electronic energy E_e.

The combination of a Slater determinant as the trial wavefunction (Equation 2.11) with the non-relativistic time-independent electronic Schrödinger equation (Equation 2.4) yields the Hartree−Fock[20–22] (HF) method. The physical interpretation of the HF method is that the complicated many-electron problem is treated as N independent one-electron problems in which each electron moves in an average field created by all other electrons in the system. In the HF method, the many-electron problem is reduced to solving a set of coupled one-electron eigenvalue equations:

$$\hat{f}(i)|\chi_i\rangle = \varepsilon_i|\chi_i\rangle \tag{2.16}$$

where $\hat{f}(i)$ is known as the Fock operator of the i-th electron and ε_i represents the orbital energies which are eigenvalues of the corresponding Fock operators. Equation 2.16 is derived by variationally minimizing the energy expectation value $\langle \Phi_0|\hat{H}_e|\Phi_0\rangle$ with respect to the choice of MOs using Lagrange's method of undetermined multipliers[23] to constrain the MOs to be orthonormal. The Fock operator of the i-th electron is defined as:

$$\hat{f}(i) = \hat{h}(i) + \hat{v}^{HF}(i) \tag{2.17}$$

where,

$$\hat{h}(i) = -\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M}\frac{Z_A}{r_{iA}} \tag{2.18}$$

where $\hat{h}(i)$ is the one-electron Hamiltonian, which contains the kinetic and electron-nuclear potential energy terms, similar to those in Equation 2.6. The $\hat{v}^{HF}(i)$ operator represents the interaction between electron i and all other electrons in the system. It is expressed as the difference between the Coulomb operator $\hat{J}_i$ and the exchange operator $\hat{K}_i$:

$$\hat{v}^{HF}(i) = \hat{J}_i - \hat{K}_i = \sum_{j}^{N} \hat{J}_{ij} - \hat{K}_{ij} \tag{2.19}$$

where,

$$\hat{J}_{ij}|\chi_i\rangle = \left\langle \chi_j \middle| \frac{1}{r_{12}} \middle| \chi_j \right\rangle |\chi_i\rangle \tag{2.20}$$

$$\hat{K}_{ij}|\chi_i\rangle = \left\langle \chi_j \middle| \frac{1}{r_{12}} \middle| \chi_i \right\rangle |\chi_j\rangle \tag{2.21}$$

The Coulomb term has a classical interpretation as $\hat{J}_{ij}$ incorporates the Coloumbic repulsion between electron i and all other electrons ($v_j(1) = \sum_{j\neq i}\int d\vec{R}_2 \left|\chi_j(\vec{R}_2)\right|^2 r_{12}^{-1}$). Note, however, that the Coulomb term includes the physically spurious interaction of electron i with itself. In contrast to the Coulomb term, the exchange term $\hat{K}_i$ does not have a simple classical interpretation as $\hat{K}_{ij}$ originates from the antisymmetric nature of the single determinant wavefunction, leading to an additional effective (exchange) interaction between electrons with the same spin. Since $\hat{J}_{ii} = \hat{K}_{ii}$, the exchange term exactly cancels the unphysical self-interaction of the electron, and therefore the HF method is free from (one-electron) self-interaction error (this is not always the case for other quantum mechanical methods, such as density-functional theory – see Section 2.1.6).

Equation 2.16 is an integro-differential where the Fock operator has a functional dependence, through the Coulomb and exchange operators, on the solutions $\{\chi_i\}$ of the equation itself. Therefore, the HF equations are nonlinear and need to be solved via an iterative procedure called the self-consistent field (SCF) method. In the SCF method Equation 2.16 is first solved using an initial guess for the MOs. This results in a new set of MOs, which can then be used as starting point to solve Equation 2.16. This iterative procedure is repeated until convergence is achieved, i.e., the calculated MOs and mean-field potentials no longer change, and the resulting MOs are eigenfunctions of the Fock operators.

In 1951, Roothaan[24] and Hall[25] demonstrated that Equation 2.16 can be converted into a matrix eigenvalue problem using the basis set approximation. Their pioneering work laid the foundation of modern computational chemistry as it paved a way to solve the electronic Schrödinger equation (Equation 2.4) approximately and efficiently using computer technology. In the basis set approximation, each MO is represented by a linear combination of a finite number of fixed basis functions. In molecular calculations, these basis functions are themselves linear combinations of Gaussian functions centered on the atomic positions, and are known as atomic orbitals (AOs) φ. Mathematically:

$$\chi_i = \sum_{\mu=1}^{K} C_{\mu,i}\varphi_\mu \tag{2.22}$$

where $C_{\mu,i}$ are the expansion coefficients, determined during the SCF iterative procedure, and φ_μ are the basis functions. In Equation 2.22, μ is an index that runs over the set of K basis functions. When atomic orbitals are used as basis functions, Equation 2.22 is known as the Linear Combination of Atomic Orbitals (LCAO) approach. In the basis set approximation, Equation 2.16 can be reformulated into the so-called Roothaan−Hall matrix equation:

$$\boldsymbol{FC} = \boldsymbol{SC\varepsilon} \tag{2.23}$$

where $\boldsymbol{C}$ is the $K \times K$ square matrix of the expansion coefficients $C_{\mu,i}$, $\boldsymbol{\varepsilon}$ is a diagonal matrix of the orbital energies ε_i; $\boldsymbol{F}$ is the $K \times K$ Fock matrix and $\boldsymbol{S}$ is the $K \times K$ overlap matrix:

$$F_{\mu v} = \int d\boldsymbol{r}_1 \varphi_\mu^* \hat{f} \varphi_v \qquad (2.24)$$

$$S_{\mu v} = \int d\boldsymbol{r}_1 \varphi_\mu^* \varphi_v \qquad (2.25)$$

The $\boldsymbol{S}$ matrix is not an identity matrix because the basis functions are generally normalized but not orthogonal to each other. If $\boldsymbol{S}$ were the identity matrix, then Equation 2.23 would just have the simple form of a usual matrix eigenvalue problem $\boldsymbol{FC} = \boldsymbol{C\varepsilon}$, which is easier to solve than Equation 2.23. In practice, this transformation is achieved by orthonormalization of the initial basis set. The final form of Equation 2.23 after orthogonalization of the basis reduces to a diagonalization problem with the $\boldsymbol{F}$ matrix. Unfortunately, this must be done iteratively, as mentioned earlier, because the matrix elements of $\boldsymbol{F}$ depend on the solution of this eigenvalue equation.

2.1.4 Basis Sets and Effective-core Potentials

Basis Sets

In order to apply the basis set approximation, we need to specify the set of basis functions used to describe the MOs. We now discuss the various types of basis sets[26–28] available for use in quantum chemistry. As mentioned earlier, a basis set refers to a collection of pre-defined one-electron basis functions (atomic orbitals) whose linear combination yields the desired MOs for the trial wavefunction. When calculating molecular properties, a careful selection of the basis set is required because the quality of the wavefunction and the accuracy of properties derived from it depends on the basis functions of the chosen basis set. It should be noted that using too many basis functions increases the computational cost of the calculation.

For a system of N electrons, the solution of the HF eigenvalue problem in Equation 2.16 (or Equation 2.23, in practice) yields a set of $\{\chi_i\}$ orthonormal MOs with orbital energies $\{\varepsilon_i\}$**.** The N MOs with the lowest energies are called the occupied orbitals. (In the spin-restriced HF formalism, where pairs of orbitals share the same spatial part, each orbital can accommodate two electrons so the $N/2$ lowest energy orbitals are occupied.) The Slater determinant composed of these MOs is the HF ground-state wavefunction. It is the best variational approximation to the ground-state of the system for a single determinant form. The remaining MOs of the set $\{\chi_i\}$ are called virtual or unoccupied orbitals. In principle, an infinite number of basis functions would yield infinite number of MOs as solutions to the HF equation. Due to computational limitations, only a finite set of basis functions can be used to solve the HF equations. A finite basis set containing K basis functions leads to a set of $2K$ MOs (K spin-up and K spin-down orbitals). Of these MOs, N are occupied orbitals and $2K - N$ are unoccupied orbitals. The quality of the final MOs depends on the size of the basis set. In the limit of infinite basis functions, known as the complete-basis-set limit (CBS), the energy expectation value $\langle \Phi_0 | \hat{H}_e | \Phi_0 \rangle$ equals the exact HF energy**.** The difference between the energy expectation value from a HF calculation with a finite basis set and the exact HF energy is referred to as the error arising from the incompleteness of the basis set.

The atomic orbitals in the basis set can have different mathematical forms. Some basis sets are composed of atom-centered basis functions written as fixed linear combinations of atomic orbitals known as primitives. A common type of function used to build the atomic orbitals is known as Slater-type orbitals (STOs). The STOs have an exponential dependence on the distance between the nucleus and electron, and resemble the exact wavefunction that emerges from the solution of the Schrodinger equation for the hydrogen atom. The angular part of the STOs can be written in terms of spherical or Cartesian coordinates:

$$\varphi_\mu^{STO} = N r^{n-1} e^{-\xi r} Y_{l,m}(\theta, \varphi) \tag{2.26}$$

$$\varphi_\mu^{STO} = N r^{n-1} e^{-\xi r} x^a y^b z^c \tag{2.27}$$

where N is a normalization constant, ξ is a constant exponent, r is the distance of the electron from the atomic nucleus, and $Y_{l,m}(\theta, \varphi)$ are real spherical harmonic functions. The n, l, and m are integer numbers related to the atomic principal, azimuthal, and magnetic quantum numbers. x, y, and z are the Cartesian coordinates, and a, b, and c are integer numbers that control the shape of the function in Cartesian form just like the spherical harmonics do in the polar form. For example, φ_μ^{STO} has s-type angular symmetry if $a + b + c = 0$, φ_μ^{STO} has p-type angular symmetry if $a + b + c = 1$, and so on.

STOs can approximate very well the long- and short-range behavior of MOs. However, calculating the integrals required for solving the Roothaan−Hall equation (Equation 2.23) is difficult with a STO basis set, since no analytical solutions to these integrals are available. The complexity of the calculation using STOs is mainly linked with the integrals that involve functions centered on different nuclei, such as the Coulomb and exchange integrals that describe electron-electron interactions. To circumvent this issue, Boys proposed an alternative to the use of STOs by changing the $e^{-\xi r}$ Slater function to a $e^{-\xi r^2}$ Gaussian function,[29] the so-called Gaussian-type orbitals (GTOs). GTOs allow the derivation of analytical solutions and recurrence relations for the integrals required in the SCF procedure. The angular part of the GTOs can also be written in terms of spherical and Cartesian coordinates***:***

$$\varphi_{\mu}^{GTO} = Nr^{2n-2-l}e^{-\xi r^2}Y_{l,m}(\theta,\varphi) \tag{2.28}$$

$$\varphi_{\mu}^{GTO} = Ne^{-\xi r^2}x^a y^b z^c \tag{2.29}$$

where ξ is the orbital exponent that controls the radial extent of the Gaussian function and the other variables have the same meaning as for the STOs. A Gaussian function with a large value of ξ does not radially extend very far from the parent atom as it decreases very quickly with increasing r. Conversely, a Gaussian function with a small value of ξ can radially extend relatively far from the nucleus.

Despite being less efficient at representing MOs than STOs, the major advantage of Gaussian functions comes from the Gaussian product theorem[8,30,31], which states that the product of two Gaussian functions located at two different centers can be expressed as a single Gaussian function centered around a different point that connects them. Application of this theorem simplifies the calculation of the challenging integrals involved in the HF equation because it allows for the derivation of analytical solutions and recurrence relations for them. Despite their advantages, GTOs, unlike STOs, cannot reproduce some of the features of the exact HF MOs such as the cusp at the origin. They also decay towards zero much more rapidly than STOs or the exact HF MOs. This means that the number of GTOs needed to obtain a solution to the HF equation with a given quality is greater than the number of STOs required. The somewhat inadequate form of GTOs can be alleviated by using fixed linear combinations of several primitive Gaussian functions as basis functions. These linear combinations are known as contractions, and they lead to contracted Gaussian-type orbitals (CGTOs):

$$\varphi_{\mu}^{CGTO} = \sum_{i=1}^{n} d_{i\mu}\varphi_i(\xi_{i\mu}) \tag{2.30}$$

where n, the number of GTOs, is referred to as the length of the contraction, $d_{i\mu}$ are the contraction coefficients, and φ_i are the GTOs with exponent $\xi_{i\mu}$. Note that in a CGTO, the contraction coefficients and exponents are pre-determined and remain constant during a calculation.

The standard basis sets used in quantum chemistry consist of basis functions built as CGTOs. Quantum chemists have devised short-hand notation to denote such standard basis sets. A shell is the combination of the principal and azimuthal quantum numbers. For instance, a carbon atom has three shells associated with its electronic configuration: 1s, 2s, and 2p. A minimal basis set contains one contraction per shell and therefore has the minimum number of functions required to accommodate all the electrons in the system. For example, hydrogen and helium atoms require a single s-type symmetry function (1s), lithium to neon atoms require two *s*-type symmetry functions (1s, 2s) and one additional *p*-type symmetry function (2p) for carbon to neon, and so on. Minimal basis sets are too small to represent MOs properly as one contraction per shell does not allow for the accurate description of the bonding of an atom in different molecules. Using more than one contraction per shell provides more basis functions to represent MOs than a minimal basis set. A basis set employing two contractions per shell is called a double-zeta (double-ζ or 2-ζ) basis set. Similarly, triple-zeta (triple-ζ or 3-ζ), quadruple-zeta (quadruple-ζ or 4-ζ), and quintuple-zeta (quintuple-ζ or 5-ζ) basis sets respectively employ three, four, and five contractions per shell, respectively. In general, an *N*-*ζ* basis set employs *N* contractions per shell. An alternative approach is to use multiple contractions for the atom's valence shells while using one or a few contractions for the core shells. Basis sets that use this approach are known as split-valence basis sets. The rationale behind split-valence basis sets is that valence electrons are actively involved in forming chemical bonds, unlike the core electrons, therefore more flexibility is required to treat the former than the latter.[32–34] Constructing basis sets with this approach reduces the computational cost while still accurately capturing the chemistry of the system.

In addition to having multiple contractions per atomic shell, additional basis functions are often added to increase the flexibility of the basis set. These basis functions are known as either polarization functions or diffuse functions and serve different purposes. Polarization functions are basis functions with higher angular momentum quantum number than the highest occupied atomic shell in the ground-state configuration of the corresponding atom. For example, let us take the case of a hydrogen atom. The electron cloud in an isolated hydrogen atom in its ground state has spherical symmetry. However, when the hydrogen atom is present in a molecule, the electron cloud is distorted due to the presence of the other nuclei. A *p*-type function can be added to a hydrogen atom's basis set to allow the hydrogen electron density to break spherical symmetry. Similarly, underlying basis sets of the first- and second-row elements can incorporate *d*-type functions to allow for polarization. On the other hand, diffuse functions are basis functions that allow for the accurate modeling of chemical species that have a significant amount of electron density located far from the nuclear centers, such as anions. Diffuse functions are Gaussian functions with small exponent values that radially extend further into space than the other functions in the basis set.

A large variety of basis sets are available for use in quantum chemistry. Most software packages have a built-in library of standard basis sets. One can also download these standard basis sets from the EMSL Basis Set Exchange[35,36]. The next few paragraphs briefly describe the four family of basis sets most commonly used in quantum chemistry.

Pople-style basis sets: The series of basis sets designed by Pople and coworkers are perhaps the most extensively used in quantum chemistry.[37–39] The minimal basis sets are denoted as $STO\text{-}NG$, where N refers to the number of GTOs entering a contraction (see Equation 2.30). In addition, the double-zeta and triple-zeta split-valence basis sets in the Pople family are denoted $n\text{-}ijG$ and $n\text{-}ijkG$, respectively. The n in $n\text{-}ijG$ or $n\text{-}ijkG$ refers to the number of GTOs in the single contraction that describes the core shells while i, j, and k after the hyphen represent the number of GTOs entering each contraction for the different valence shells – it is a valence-double-zeta or valence-triple-zeta basis set depending on whether there are two or three such numbers, respectively. Polarization functions are denoted via asterisks or sometimes by the specific angular momentum label and the number of additional functions. Diffuse functions are represented by either a single or double "+" sign, indicating an extra *s*- and *p*-type diffuse function for any atom between helium to calcium (single "+") and an additional *s*-type diffuse function for hydrogen (double "+"). For example, the STO-3G basis set is a minimal basis set in this family, with 3 GTOs entering a single contracted function per atomic shell. The 6-31G(d) (also called 6-31G*) basis set is a double-zeta split-valence basis set with 6 GTOs and a single contraction for the core shells, 3 GTOs for the first set of valence AOs, and 1 GTO for the second set of valence AOs, plus one *d*-type polarization function for non-hydrogen atoms. The 6-311++G(d,p) (also called 6-311++G**) basis set is a triple-zeta split-valence basis set with 6 GTOs forming a single contraction for each core shell, 3 GTOs for the first set of valence AOs, 1 GTO for both the second and third set of valence AOs, one *d*-type polarization function for non-hydrogen atoms (first asterisk), a single *p*-type polarization function for hydrogen atoms (second asterisk), a set of *s*-type and *p*-type diffuse functions for heavy atoms (first "+"), and one *s*-type diffuse function for hydrogen atoms (second "+").

Correlation consistent basis sets: These basis sets developed by Dunning and coworkers are designed for accurate calculations using post-Hartree−Fock methods (see Section 2.1.5).[40–42] This class of basis sets is denoted $cc\text{-}VNZ$, where cc stands for correlation consistent, N is the cardinal number of the basis set (2 or D for double-zeta, 3 or T for triple-zeta, 4 or Q for quadruple-zeta, and so on), and VZ represent the split-valence zeta nature of the basis set. Basis sets in this family usually have a higher number of basis functions per atom than the Pople basis sets, making them more expensive to use (but more accurate). When polarization functions are added to the basis sets, the representation changes to $cc\text{-}pVNZ$. The correlation consistent basis sets can be augmented further with diffuse functions and are denoted by the prefix "aug-". A commonly used basis set is $aug\text{-}cc\text{-}pVTZ$, a triple-zeta split-valence basis set with polarization and diffuse functions.

Karlsruhe basis sets: Ahlrich and coworkers developed a class of basis sets referred to as the Karlsruhe basis sets.[43–45] The latest generation of these basis sets is denoted by the prefix "$Def2$-", and they are the default basis set in the Turbomole[46] software package. The Karlsruhe basis sets were designed to be used with density functional theory (see Section 2.1.6). When the "$Def2$-" is followed by SV or NZV, the basis set is a split valence basis set or N-zeta valence basis set, respectively. Additional polarization and diffuse functions are specified with a suffix P and D, respectively. A commonly used basis set of this family is the $Def2$-$TZVPD$, a triple-zeta valence basis set with additional polarization and diffuse functions.

Polarization consistent basis sets: Jensen and coworkers developed the family of basis sets known as the polarization consistent basis sets.[47–50] These basis sets employ more polarization functions, compared to the other families, to produce faster convergence to the HF limit than might be achieved using correlation-consistent basis sets. These basis sets were optimized for use with HF and density functional theory methods. The nomenclature for these basis sets is pc-X, where X represents the cardinal number of the basis set minus one. Additional diffuse functions can be specified by the "aug-" prefix, as in the correlation consistent basis sets. A variant of the initially developed pc-X also exists which is known as $pcseg$-X.[51] The seg refers to a change in the contraction scheme when building contracted GTOs from primitive Gaussian functions.

Effective-core Potentials

For heavy elements of the periodic table, the core electrons experience substantial relativistic effects as a consequence of large nuclear charges.[52,53] Neglecting these effects in computational modeling can lead to significant errors in properties of heavy elements.[54–56] Since relativistic effects are not accounted for by the use of non-relativistic Schrödinger equation, alternative approaches have to be adopted for heavier elements. In addition, due to the large number of electrons, heavy elements require more basis functions to expand the corresponding MOs, thereby leading to a significantly increased computational expense. For instance, the 3-21G split-valence double-zeta basis set employs only 9 basis functions for a single carbon atom but more than three times the number of basis functions (33) for a single antimony atom directly three rows below carbon.

One efficient approach to address both difficulties related to heavy elements, the large number of basis functions required and the relativistic effects, is to eliminate a number of core electrons from calculations by use of atom-centered potential functions. These atom-centered potential functions are referred to as effective-core potentials[57–60] (ECPs). Construction of ECPs requires proper description of the influence of eliminated core electrons on the valence electrons and further involves the inclusion of scalar relativistic effects by a suitable parametrization to relativistic reference data. Therefore, by using ECPs, the electrons and their corresponding basis functions are removed from the calculation leading to computational savings and also relativistic effects are incorporated into what would otherwise be a non-relativistic Schrödinger equation.[61–63] ECPs are generally written as (fixed) linear combination of Gaussian-type functions, since this greatly simplifies the calculation of the associated molecular integrals:

$$V_l(r) = \sum_{k=1}^{N} c_{l,k} r^{n_{l,k}-2} exp(-\xi_{l,k} r^2) \tag{2.31}$$

In Equation 2.31, r is the distance from the nucleus, l is the angular momentum quantum number, $n_{l,k}$ is the power of the radial pre-factor, generally set to 2, $c_{l,k}$ is the contraction coefficient, and $\xi_{l,k}$ is the Gaussian exponent. Generally, the potential with the highest angular momentum ($LMAX$) is parametrized first, and then lower angular momentum potentials are parametrized with the analytical potential for $LMAX$ subtracted. Following the early works of Goddard[64] and Kahn and co-workers[65,66], the general form in which ECPs are written is:

$$\hat{V}_{ECP} = V_{LMAX}(r) + \sum_{l=0}^{LMAX} \sum_{m=-l}^{l} [\, V_l(r) - V_{LMAX}(r)] |Y_{lm} >< Y_{lm}| \tag{2.32}$$

where l is the angular momentum quantum number ranging from 0 to $LMAX$, m is the magnetic quantum number associated with l, and Y_{lm} are real spherical harmonic functions. ECPs given by Equation 2.32 are centered on each atom to which they are applied and added to the one-electron Hamiltonian (Equation 2.18):

$$\hat{h}(i) = -\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M} \left[\frac{Z_A^{eff}}{|\, \vec{r}_i - \vec{R}_A \,|} - \hat{V}_{ECP}^{A}(r) \right] \tag{2.33}$$

where i is an electron, A is a nucleus, Z_A^{eff} is the effective nuclear charge (the atomic number minus the number of core electrons replaced by the ECP), and the potential $\hat{V}_{ECP}^{A}(r)$ is the ECP centered on A. There are two kinds of integrals that result from Equation 2.32: Type II corresponding to the integrals involving projections with real spherical harmonics due to the second "non-local" term, and Type I corresponding to the unprojected integrals due to the first "local" term. Several methods have been developed for the efficient evaluation of these integrals and the interested reader is referred to the recent work by Shaw *et al.*[67] and Mckenzie *et al.*[68] for more details and the relevant references therein.

2.1.5 Post-Hartree–Fock Approaches

Electron Correlation

Once the self-consistent MOs are obtained at the end of the SCF procedure, the HF energy (E^{HF}) is calculated:

$$E^{HF} = \sum_{i}^{N} \langle \varphi_i | \hat{h}_i | \varphi_i \rangle + \sum_{i}^{N-1} \sum_{j>i}^{N} \langle \varphi_i | \hat{v}^{HF}(i) | \varphi_i \rangle \qquad (2.34)$$

The HF approximation is the simplest, fully quantum mechanical method. Because HF neglects the correlated motion of electrons, even in the CBS limit, HF cannot yield the exact energy of a multi-electron system. The difference between the exact electronic energy (E_{exact}) and the HF energy at the CBS limit (E_{CBS}^{HF}) is called the correlation energy (E_{corr}):

$$E_{corr} = E_{exact} - E_{CBS}^{HF} \qquad (2.35)$$

Interestingly, the HF approach recovers more than 99% of the exact electronic energy, and one might assume it to be satisfactory. The remaining 1% of the exact electronic energy is the missing correlation energy. This 1% is critical to achieving the accuracy necessary for studying most problems in quantum chemistry. Consequently, a large amount of theoretical work has focused on developing new methods and techniques that approximate the correlation energy as accurately and efficiently as possible.

By definition (Equation 2.35), HF lacks all electron correlation and the missing correlation energy is classified in two ways. The dynamical correlation corresponds to the missing electron correlation effects from the HF in the context of a single determinant ground state wavefunction. For example, the correlation in the H_2 molecule is dynamical at the equilibrium interatomic bond distance. In comparison, the other correlation is known as non-dynamical or static correlation. The static correlation corresponds to the electron correlation effects that arise due to a non-single determinant form of the ground state wavefunction. For example, the correlation in infinitely stretched H_2 is static. Static correlation is also sometimes called a near-degeneracy effect, as it becomes important for systems where different electronic configurations have similar energies. For instance, trimethylenemethane biradical (or 2-

methylidenepropane-1,3-diyl) in its closed-shell singlet state yields two electronic configurations that have the same energies. Therefore, the ground state wavefunction of trimethylenemethane biradical requires to be represented via a linear combination of two Slater determinants representing the two degenerate electronic configurations. Despite the classification of correlation effects that allows a more conceptual way of thinking about them, there is no known way of separating the two types of correlation. For methods that can account for static correlation effects, see references 69–73 and references therein. In the following, a few important post-HF approaches used to account for the dynamical correlation are discussed, including the configuration interaction methods[74], coupled-cluster theory methods[75–77], and Møller–Plesset perturbation theory methods[76,78]. The starting point for all these methods is the HF single-determinant wavefunction.

Excited Determinants

The Slater determinant in the HF method is the best (in a variational sense) approximate single-determinant wavefunction for a molecule's ground state. The HF approach yields a set of $2K$ spin orbitals ($\{\chi_i\}$), where K is the number of basis functions. For an N-electron system, only N spin orbitals are occupied and $2K - N$ spin orbitals are left unoccupied. As a result, many different determinants could be formed by replacing some of the occupied orbitals in the HF determinant by unoccupied orbitals. These determinants yield higher expectation values for the energy, and they are called excited determinants. The excited determinants are created by removing electrons from occupied orbitals and placing them in unoccupied orbitals, thus changing the electronic configuration from the ground state to an excited state. In singly excited determinants, one electron occupying χ_a in the HF (reference) determinant is promoted to a virtual χ_r:

$$|\Psi_a^r\rangle = |\chi_1\chi_2 \cdots \chi_r\chi_b \cdots \chi_N\rangle \tag{2.36}$$

Similarly, in doubly excited determinants, electrons from two occupied orbitals χ_a and χ_b are promoted to the virtual orbitals χ_r and χ_s:

$$|\Psi_{ab}^{rs}\rangle = |\chi_1\chi_2 \cdots \chi_r\chi_s \cdots \chi_N\rangle \tag{2.37}$$

The total number of possible determinants formed by exciting electrons from occupied N spin orbitals to $2K - N$ virtual spin orbitals is the binomial coefficient $\binom{2K}{N} = \frac{(2K)!}{N!(2K-N)!}$. It is possible to show that an arbitrary N-electron wavefunction can be exactly expanded in terms of all unique determinants formed using excitations of the HF reference determinant from a complete set of one-electron MOs.

Therefore, the excited determinants can be used as a basis for expanding the N-electron wavefunction, in the same way that a collection of Gaussian functions was used to expand the HF MOs. If $|\Psi_0\rangle$ is assumed to be the reference HF determinant, one can write the exact N-electron wavefunction $|\Phi\rangle$ as a linear combination of all possible Slater determinants formed from a complete set of spin orbitals $\{\chi_i\}$:

$$|\Phi\rangle = c_0|\Psi_0\rangle + \underbrace{\sum_{ra} c_a^r|\Psi_a^r\rangle}_{Singles} + \underbrace{\sum_{\substack{a<b\\ r<s}} c_{ab}^{rs}|\Psi_{ab}^{rs}\rangle}_{Doubles} + \underbrace{\sum_{\substack{a<b<c\\ r<s<t}} c_{abc}^{rst}|\Psi_{abc}^{rst}\rangle}_{Triples} + \cdots \qquad (2.38)$$

where $a < b$ and $r < s$ refers to summing over the unique pairs of occupied and virtual spin orbitals, respectively. c_0, c_a^r, c_{ab}^{rs}, c_{abc}^{rst},… are the various expansion coefficients associated with the various Slater determinants $|\Psi_0\rangle$, $|\Psi_a^r\rangle$, $|\Psi_{ab}^{rs}\rangle$, $|\Psi_{abc}^{rst}\rangle$,…. respectively. The infinite set of N-electron determinants $\{|\Psi_i\rangle\} = \{|\Psi_0\rangle, |\Psi_a^r\rangle, |\Psi_{ab}^{rs}\rangle, |\Psi_{abc}^{rst}\rangle, \ldots\}$ is a complete set for the expansion of any N-electron wavefunction.

Configuration Interaction

A linear variational method can be used to determine the expansion coefficients of the best approximate wavefunction and the approximate electronic energy of a trial wavefunction obtained by truncation of the Slater determinant expansion in Equation 2.38. Since the trial wavefunction is written as a linear combination of Slater determinant, application of the variational method results in a matrix eigenvalue problem involving the Hamiltonian matrix, $H_{ij} = \langle\Psi_i|\hat{H}_e|\Psi_j\rangle$.

This procedure is known as the configuration interaction (CI) method. In the limit of a complete set of determinants $\{|\Psi_i\rangle\}$ in the trial wavefunction, this gives the exact solution provided the one-electron basis for the spin orbitals is complete. Unfortunately, this method cannot be implemented because a matrix with infinite elements cannot be used in practice. However, if a finite set of spin orbitals is used, $\binom{2K}{N}$ determinants can be formed, although they do not lead to a complete N-electron basis. The CI procedure using all determinants that can be built from a finite basis set is referred to as full CI. Because the number of determinants increases factorially with the system size, full CI is only practically possible for very small molecules.

The alternative to the full CI procedure is truncated CI methods like CIS, CID, CISD, CISDT,…. In these methods, Equation 2.38 is truncated and the trial wavefunction contains only specific excitations like singles, doubles, singles and doubles, singles with doubles and triples, etc. The CI methods resulting from truncated wavefunction scale less favorably than HF with system size and, most importantly, are not “size consistent”. A method is size consistent if the energy of a pair of molecules (A and B) at infinite separation equals the sum of the energies of the two molecules calculated independently:

$$E^{AB}(r \to \infty) = E^{A} + E^{B} \tag{2.39}$$

Coupled-cluster Theory

The exact N-electron wavefunction in Equation 2.38 can also be written in a simplified form as:

$$|\Phi\rangle = (1 + \sum_{i=1}^{N} \hat{C}_i)|\Psi_0\rangle \tag{2.40}$$

where $\hat{C}_i$ are the operators that represent the i-fold excitations that create all possible excited determinants of order i. The coupled-cluster (CC) method is based on using a trial wavefunction in which $(1 + \sum_{i=1}^{N} \hat{C}_i)$ from Equation 2.40 is replaced by an exponential operator $e^{\hat{T}}$:

$$|\Phi\rangle = e^{\hat{T}}|\Psi_0\rangle \tag{2.41}$$

where $\hat{T}$ is the cluster operator, defined by i-fold electron excitation operator $\hat{T}_i$ as:

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \cdots \tag{2.42}$$

Equation 2.42 can be truncated to eliminate higher-order excitations. The exponential operator in Equation 2.41 is then expanded in a Taylor series of $\hat{T}$. This creates a truncated expression of the infinite CI expansion (Equation 2.38) that, unlike in CI methods, preserves the self-consistency of the resulting method. For example, if Equation 2.42 is truncated to include only single and doubles excitations, the $\hat{T}$ operator is equal to $\hat{T}_1 + \hat{T}_2$ and Equation 2.41 simplifies to:

$$|\Phi\rangle = e^{\hat{T}_1 + \hat{T}_2}|\Psi_0\rangle \tag{2.43}$$

Expanding the exponential into its Taylor series gives:

$$|\Phi\rangle = (1 + \hat{T}_1 + \hat{T}_2 + \frac{1}{2!}\hat{T}_1^2 + \hat{T}_1\hat{T}_2 + \frac{1}{2!}\hat{T}_2^2 + \cdots)|\Psi_0\rangle \tag{2.44}$$

In the limit where all possible excitations are included, CC theory is equivalent to full CI. However, truncated CC methods have a significant advantage over truncated CI methods in that they are size consistent. For example, in CI and CC with single and double excitations (CISD and CCSD methods), the CISD wavefunction only includes terms associated with $\hat{T}_1$, $\hat{T}_2$, and $\hat{T}_1^2$. In contrast, CCSD includes terms associated with $\hat{T}_1$, $\hat{T}_2$, $\hat{T}_1^2$, $\hat{T}_1\hat{T}_2$, $\hat{T}_2^2$,… as given by Equation 2.44. The CISD truncated expansion only contains singly and doubly excited determinants, whereas the presence of operators like $\hat{T}_1^2$, $\hat{T}_1\hat{T}_2$, $\hat{T}_2^2$,… in the CCSD expansion implicitly generates doubly, triply, quadruply, and other higher-order excited determinants. The inclusion of these additional terms in the wavefunction generated by using the exponential of the cluster operator ensures the size-consistency of CCSD, unlike CISD. The computational cost of CCSD and higher-order truncated CC methods scales quickly with the system size, and these methods can be applied to systems with only a few tens of atoms. A variant of the CCSD method known as CCSD(T) also exists where the (T) refers to the indirect inclusion of triple excitations via perturbation theory.[79,80] The CPU time of the CCSD(T) method scales as N^7 (where N represents the system size). For comparison, the HF method scales somewhere between N^3 to N^4. When combined with large basis sets, CCSD(T) is known to reproduce various experimental properties with very high accuracy, such as bond lengths within 0.1−0.2 pm, bond angles within 0.1−0.2°, dipole moments within 0.01−0.02 D, atomization energies within 0.25−0.50 kcal/mol, reaction enthalpies within 0.50 kcal/mol, conformational barriers within 0.25 kcal/mol, and non-covalent interaction energies within 0.25 kcal/mol.[81]

Møller−Plesset Perturbation Theory

Another post-HF approach discussed in this section is the Møller−Plesset perturbation theory (or MPn methods). MP perturbation theory is based on applying Rayleigh−Schrödinger perturbation theory on the optimized HF wavefunction Φ_0. In many-body perturbation theory, the Hamiltonian of a perturbed system ($\hat{H}_P$) is written as:

$$\hat{H}_P = \hat{H}_0 + \lambda\hat{V} \tag{2.45}$$

where $\hat{H}_0$ is the unperturbed Hamiltonian, which corresponds to a simpler problem and has known solutions. $\hat{V}$ is assumed to be a small perturbation, and λ is a parameter (varying between 0 and 1) determining the perturbation's strength. Through the process of expressing $\hat{H}_P$ as shown in Equation 2.45, a systematic procedure to improve the solutions associated with $\hat{H}_0$ can be devised so that they become closer and closer to the exact solutions of the $\hat{H}_P$. The perturbed Schrödinger equation can be written as:

$$\hat{H}_P \Psi_P = E_P \Psi_P \tag{2.46}$$

where the perturbed wavefunction and energy associated with Equation 2.46 are expressed as a power series in λ:

$$\Psi_P = \sum_{i=0}^{\infty} \lambda^i \Psi^{(i)} \tag{2.47}$$

$$E_P = \sum_{i=0}^{\infty} \lambda^i E^{(i)} \tag{2.48}$$

Methods based on the MP perturbation theory estimate the correlation energy by applying many-body perturbation theory. In MP perturbation theory, the unperturbed Hamiltonian $\hat{H}_0$ is the sum over the Fock operators (see Equations 2.17 and 2.18). The zeroth-order energy is then the sum of the energy of the MOs. This leads to double counting of electron-electron repulsion for each pair of electrons; thus, the perturbation is the exact electron repulsion operator minus twice the average electron repulsion operator:

$$\hat{V} = \hat{V}_{ee} - 2\langle \hat{V}_{ee} \rangle \tag{2.49}$$

The first-order energy correction is the average of the first-order perturbation operator over the zeroth-order wavefunction (Φ_0)***:***

$$E^{(1)} = \left\langle \Phi_0 \middle| \hat{V} \middle| \Phi_0 \right\rangle \tag{2.50}$$

The total energy through the first-order is then:

$$E(MP1) = E^{(0)} + E^{(1)} = \left\langle \Phi_0 \middle| \hat{H}_0 + \hat{V} \middle| \Phi_0 \right\rangle = E_{HF} \tag{2.51}$$

Equation 2.51 gives the same result as the HF method. As the MP1 energy is equal to the HF energy, account of electron correlation begins with the second-order energy correction (MP2 or higher methods):

$$E(MP2) = E_{HF} + E^{(2)} \tag{2.52}$$

The second-order energy correction ($E^{(2)}$) is given as:

$$E^{(2)} = -\frac{1}{4} \sum_{ij}^{occ} \sum_{ab}^{virt} \frac{|\langle ij \,||\, ab \rangle|^2}{\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j} \tag{2.53}$$

$$\langle ij \,||\, ab\rangle = \int d\boldsymbol{x}_1 d\boldsymbol{x}_2 \chi_i^*(\boldsymbol{x}_1)\chi_j^*(\boldsymbol{x}_2) r_{12}^{-1}(1-\hat{P}_{12})\chi_a(\boldsymbol{x}_1)\chi_b(\boldsymbol{x}_2) \tag{2.54}$$

The i, j and a, b in Equation 2.53 are occupied and virtual orbitals, respectively. The ϵ_a, ϵ_b, ϵ_i, ϵ_j are the HF orbital energies associated with occupied and virtual orbitals. $\hat{P}_{12}$ is a permutation operator that exchanges the coordinates of electrons 1 and 2. The higher order MP*n* methods that include additional perturbative energy corrections are referred to as MP3, MP4, and so on. Among the MP based methods, MP2 is the most popular in quantum chemistry due to its relatively modest N^5 scaling with system size and due to MP2 being able to recover a significant amount of the missing correlation energy. As the MP*n* methodology is not variational, it is possible that the estimates for the correlation energy can lead to an overestimation of the total energy when compared to the exact energy. In addition, there is no guarantee that further perturbative corrections beyond MP2 (like in MP3 and higher) would converge smoothly to an asymptote particularly when a finite basis set is used.[82]

Alternatives to Conventional Post-HF Approaches

Since calculations of various molecular properties with large basis sets and the CCSD(T) method are practically feasible only for very small systems, various basis-set extrapolation[83] techniques have been developed to approximate the results of CBS limit CCSD(T) (CCSD(T)/CBS). In the computational chemistry community, CCSD(T)/CBS is also commonly referred to as the "gold standard" approach and is often used to obtain benchmark quality reference data for thermochemical quantities and intermolecular interactions.[84] In extrapolation methods that attempt to estimate CCSD(T)/CBS, calculations for the HF energies and correlation energies are carried out separately using multiple basis sets of increasing size. The resulting energies are estimated using a three-parameter[85,86] or two-parameter[87,88] extrapolation formulas. An example of one such extrapolation approach is given in Chapter 6.

Because converging the CC expansion with respect to the number of excitations and the one-electron basis set size is so expensive, an active area of research is the development of methods that allow accurate CC calculations at modest computational cost. We briefly describe the "explicitly correlated" and "local correlation" methods.[89–93] Explicitly correlated approaches introduce additional functions that depend explicitly on the inter-electronic distance into the multi-electron wavefunction expansion. This approach offers an efficient way of accounting for dynamical electron correlation effects with reduced size basis sets. For example, explicitly correlated CCSD with a 3-ζ basis set can attain the same accuracy as conventional CCSD with a 5-ζ basis set.[94,95] Whereas, local correlation methods offer a way to lower the computational cost of post-HF methods by exploiting the locality or short-range character of dynamical correlation between pairs of electrons. In local correlation methods, the reduction in computational cost is mainly achieved by the use of localized orbitals[96] and selection of important electron pairs based on their inter-electronic distances. In recent years, local correlation methods that also utilize explicitly correlated approaches have been developed.[92,93] A local correlation method called domain based local pair natural orbital CCSD(T) (or DLPNO-CCSD(T)[97–99] method) has been used in place of conventional CCSD(T) when dealing with moderate-sized systems in Chapter 6.

In the past decades, various multistep theoretical procedures referred to as "composite methods"[100] have also been developed to obtain thermochemical and kinetic properties with sub-kcal/mol accuracy compared to experimental results. In general, composite methods combine calculations performed with series of computationally lower-cost methods to approximate CCSD(T)/CBS or complete-basis-set-limit full CI results. Many composite methods also include additional corrections to account for various missing contributions from the total energy obtained from solving the non-relativistic Born−Oppenheimer Schrödinger equation like, for instance, core-valence, relativistic, spin-orbital, Born−Oppenheimer, and zero-point vibrational energy corrections. An example of a composite method used in this thesis work is presented in Chapter 5.

2.1.6 Density Functional Theory

An alternative way to incorporate electron correlation is via density functional theory[101–104] (DFT). The essential quantity in DFT is not the wavefunction but the electron density $\rho(r)$, a three-dimensional scalar function that describes the probability of finding electrons in real space:

$$\rho(r) = \sum_i |\chi_i(\boldsymbol{r})|^2 \tag{2.55}$$

where χ_i are occupied molecular orbitals. In the past decades, DFT has become one of the most popular electronic structure methods, primarily because it offers a good compromise between computational efficiency and accuracy. DFT relies on the two Hohenberg−Kohn theorems[105]. The first Hohenberg−Kohn theorem establishes that there is a one-to-one correspondence between the electron density and the many-electron wavefunction. Therefore, the energy, the wavefunction, and any observable of a system can be entirely determined by its electron density, unlike the full CI method, where properties derived from the exact wavefunction at first require the generation of all excited determinants from the reference HF determinant. The second Hohenberg−Kohn theorem is similar to the variational principle discussed earlier. It states that for any valid trial electron density $\tilde{\rho}$, $E[\tilde{\rho}(r)]$ is higher than the exact ground state energy. In principle, DFT is an exact theory. However, the Hohenberg−Kohn theorems do not provide a recipe for determining the mathematical form of the exact energy functional. The pursuit of finding improved approximation for the energy functional of the density has been a driving force behind many developments in DFT.

As the exact energy functional is not known, the approximation to $E[\rho(r)]$ used in DFT methods is what determines their accuracy. $E[\rho(r)]$ can be divided into separate functionals that represent different contributions:

$$E[\rho(r)] = T_e[\rho(r)] + V_{en}[\rho(r)] + V_{ee}[\rho(r)] \tag{2.56}$$

$$V_{ee}[\rho(r)] = J[\rho(r)] + K[\rho(r)] \tag{2.57}$$

In Equation 2.56, $T_e[\rho(r)]$ yields the contributions for the kinetic energy of electrons, $V_{en}[\rho(r)]$ is called the "external potential" and yields the contributions for electron-nuclei attraction, and $V_{ee}[\rho(r)]$ yields the contributions for electron-electron repulsion. Equation 2.57 shows that $V_{ee}[\rho(r)]$ could be further split into contributions for a Coulomb part $J[\rho(r)]$ and an exchange part $K[\rho(r)]$. The functionals $V_{en}[\rho(r)]$ and $J[\rho(r)]$ depend directly on the electron density and can be calculated straightforwardly:

$$V_{en}[\rho(r)] = -\sum_{A=1}^{M} \int \frac{Z_A \rho(\boldsymbol{r})}{|\boldsymbol{R_A} - \boldsymbol{r}|} d\boldsymbol{r} \tag{2.58}$$

$$J[\rho(r)] = \frac{1}{2} \int \frac{\rho(\boldsymbol{r})\rho(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} d\boldsymbol{r} d\boldsymbol{r}' \tag{2.59}$$

Based on the uniform electron gas (UEG) model, the expression for $T_e[\rho(r)]$ was first approximated by Thomas[106] and Fermi[107] while that of $K[\rho(r)]$ was done by Dirac[108] as follows:

$$T_e^{UEG}[\rho(r)] = \frac{3}{10}(3\pi)^{2/3} \int \rho(\boldsymbol{r})^{5/3} d\boldsymbol{r} \tag{2.60}$$

$$K^{UEG}[\rho(r)] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} \int \rho(\boldsymbol{r})^{4/3}\, d\boldsymbol{r} \tag{2.61}$$

The above equations were early attempts to approximate the true energy functional and did not yield valuable results for chemical applications. It was not until the formulation of Kohn−Sham DFT that the theory began gaining popularity.[109] In the Kohn−Sham approach to DFT (KS-DFT), orbitals φ are introduced to evaluate the kinetic energy represented as $T_e^{KS}[\varphi]$ via a fictitious reference system of non-interacting quasi-particles that is constrained to have the same density as the real system. Using this approach, most of the kinetic energy can be recovered, and the missing kinetic energy, as well as the overall correlation and the exchange effects are described by the so-called exchange-correlation functional E_{XC}. E_{XC} is usually divided into the exchange and correlation parts represented by E_X and E_C, respectively. The total DFT energy is therefore given as:

$$E^{DFT} = T_e^{KS}[\varphi] + V_{en}[\rho(r)] + J[\rho(r)] + E_{XC}[\rho(r)] \tag{2.62}$$

$$E_{XC}[\rho(r)] = E_X[\rho(r)] + E_C[\rho(r)] \tag{2.63}$$

$$T_e^{KS}[\varphi] = -\frac{1}{2}\sum_{i=1}^{N}\langle\varphi_i|\hat{\nabla}_i^2|\varphi_i\rangle \tag{2.64}$$

The exchange energy term $E_X[\rho(r)]$ in Equation 2.63 is defined as the difference between the expectation value of the many-body electron-electron energy term **(**V_{ee}) and the classical electron-electron repulsion term:

$$E_X[\rho(r)] = \langle\sum_{i>j}\frac{1}{r_{ij}}\rangle - J[\rho(r)] \tag{2.65}$$

Equation 2.65 shows that $E_X[\rho(r)]$ is an energetic contribution that corrects for the double-counting of electrons in $J[\rho(r)]$. In a one-electron system, the exchange term should cancel exactly the spurious self-interaction of the electron with itself coming from $J[\rho(r)]$, like the exchange term in the HF method does, but in most DFT methods based on approximate exchange functionals, it does not. The correlation energy term $E_C[\rho(r)]$ in Equation 2.63 is defined as the missing energy contributions necessary to make $E_{XC}[\rho(r)]$ exact.

KS-DFT is nowadays the most commonly used DFT method, and therefore, the prefix KS is usually dropped for simplicity. The electronic energy and the Kohn-Sham orbitals are obtained iteratively by solving the Kohn−Sham equations,

$$\hat{f}_i^{KS} = \left[\hat{h}_i[\rho(r)] + \sum_j\left(\hat{J}_{ij}[\rho(r)] + V_{XC}[\rho(r)]\right)\right]\varphi_i = \varepsilon_i\varphi_i \tag{2.66}$$

where $\hat{f}_i^{KS}$ is a one-electron operator analogous to the Fock-operator in the HF method. The solution of the SCF problem is very similar in DFT as in HF and, therefore, the computational cost and scaling of both methods is similar. The main difference between HF and DFT is that, instead of the HF exchange operator, DFT uses an exchange-correlation potential $V_{XC}[\rho(r)]$ which is the functional derivative of the $E_{XC}[\rho(r)]$

with respect to the density. While DFT is formally an exact theory, various approximations are used to estimate the form of the unknown exchange-correlation functional. The use of these density functional approximations (DFAs) leads to deviations from the exact result. Unlike wavefunction based methods, which can be improved by the inclusion of more excited daterminants, there is no systematic way to improve DFAs. The one significant advantage of DFT over HF is that at a similar or slightly higher computational cost, DFT can include correlation effects when a sufficiently accurate approximation for $E_{XC}[\rho(r)]$ is used.

DFAs can be roughly categorized in the rungs of a hierarchical scheme referred to as the "Jacob's ladder".[110,111] DFAs that belong higher in the ladder use more complex exchange correlation energy functionals and are more accurate and expensive. However, the increase in accuracy with each rung is not guaranteed, but the general picture holds as shown in various benchmark studies.[112,113] The first rung represents the simplest approximation that is known as the local-density approximation (LDA). LDA calculates the exchange-correlation energy by assuming that the system behaves locally like a UEG. That is:

$$E_{XC} = \int \rho(\boldsymbol{r})\varepsilon_{XC}^{LDA}\big(\rho(\boldsymbol{r})\big)d\boldsymbol{r} \tag{2.67}$$

where $\varepsilon_{XC}^{LDA}\big(\rho(\boldsymbol{r})\big)$ is the exchange-correlation energy density per electron of a UEG with density ρ. The LDA exchange functional is the exchange energy of the UEG, also known as the Slater exchange energy (Equation 2.61)[108,114,115]. The LDA correlation functional is derived from Quantum Monte-Carlo calculations of the UEG[116], which were later parametrized by Vosko, Wilk, and Nusair in 1980[117] and by Perdew and Wang in 1992[118]. LDA has been widely and successfully used to describe metals, whose electronic structure is similar to that of the UEG, as well as other solids. However, LDA functionals are not practically useful for most chemical problems as they overestimate bond energies and yield poor results for thermochemistry.

The second rung of Jacob's ladder corresponds to generalized-gradient approximation (GGA) functionals, which are among some of the most popular DFAs. GGA functionals enhance LDA by using the gradient of the density $\nabla\rho(r)$ in addition to the local density to calculate the exchange correlation energy:

$$E_{XC} = \int \rho(\boldsymbol{r})\varepsilon_{XC}^{GGA}\big(\rho(\boldsymbol{r}), \nabla\rho(\boldsymbol{r})\big)d\boldsymbol{r} \tag{2.68}$$

By making the energy dependent on the density gradient, it is possible to better account for local inhomogenity in the electron density. Two popular GGA functionals are the Perdew−Burke−Ernzerhof (PBE)[119] exchange-correlation functional and the Becke 1988 (B88)[120] exchange functional with the Lee−Yang−Parr (LYP)[121] correlation functional (the combined exchange-correlation functional is known as BLYP).

The third rung of Jacob's ladder is meta-GGA functionals. Meta-GGA functionals increase the flexibility in the functional definition by using, in addition to the density and gradient, the Laplacian of the density $\nabla^2\rho(\boldsymbol{r})$ and the KS kinetic-energy density $\tau_{KS}(\boldsymbol{r})$:

$$E_{XC} = \int \rho(\boldsymbol{r})\varepsilon_{XC}^{meta-GGA}(\rho(\boldsymbol{r}), \nabla\rho(\boldsymbol{r}), \nabla^2\rho(\boldsymbol{r}), \tau_{KS}(\boldsymbol{r}))d\boldsymbol{r} \qquad (2.69)$$

where,
$$\tau_{KS}(\boldsymbol{r}) = -\frac{1}{2}\sum_{i}^{occupied} |\nabla\chi_i(\boldsymbol{r})|^2 \qquad (2.70)$$

where $\chi_i(\boldsymbol{r})$ are the self-consistently determined KS orbitals. Some of the most popular meta-GGA functionals are the Tao−Perdew−Staroverov−Scuseria (TPSS)[122] exchange-correlation functional and its revised version called revTPSS[123]. LDA, GGA and meta-GGA functionals are semi-local because they include local information like the electron density and information about its proximity via first-order or higher-order derivatives of the electron density.

The fourth rung of Jacob's ladder are the hybrid functionals that use additional information based on the occupied KS orbitals. Becke[124] showed that calculation of thermochemical properties can be significantly improved compared to semi-local functionals by using an approach where a GGA functional is mixed with a fraction of exact exchange, calculated the same way as the exchange energy in HF theory (Equarion 2.22) but using the Kohn−Sham orbitals. This approach is justified by the adiabatic connection formula[125,126], which says that the exchange-correlation energy can be represented as an integral over a coupling constant λ, whose magnitude connects the non-interacting KS system at $\lambda = 0$ to the fully interacting system at $\lambda = 1$. The general form of hybrid exchange-correlation functionals is:

$$E_{XC}^{hybrid} = E_C^{(meta-)GGA} + (1 - a_X)E_X^{(meta-)GGA} + a_X E_X^{HF} \qquad (2.71)$$

Some of the most popular hybrid functionals include the B3LYP[121,127], BHLYP[121,124], and PBE0[128]. B3LYP is a combination of Becke's 1993 exchange hybrid and LYP correlation functionals. In B3LYP, the amount of exact exchange is 20% ($a_X = 0.20$). BHLYP contains 50% of exact exchange (or $a_X =$

0.50) while PBE0 has an exact exchange fraction of 25% (or $a_X = 0.25$). Some other popular hybrid functionals based on meta-GGA functionals are from the Minnesota family of functionals like M05, M05-2X, M06, M06-2X,…(see reference 129 and references therein).

Also in the fourth rung of Jacob's ladder are the range-separated (also called long-range corrected) hybrid functionals. Similar to hybrids, range-separated hybrid functionals combine exact exchange with a semi-local functional, but they do so by partitioning the electron-electron interaction operator $1/r_{ij}$ into long-range ($\mathrm{erf}(\omega r_{ij})/r_{ij}$) and short-range ($(1 - \mathrm{erf}(\omega r_{ij}))/r_{ij}$) parts, where erf is the error function. The range of both terms is tuned by the range-separation parameter ω. The main reason to divide $1/r_{ij}$ into short- and long-range components is to recover the correct long-range behavior of the exchange-correlation potential. The exchange-correlation potential of semi-local functionals decays exponentially with distance. In the asymptotic limit, the hybrid exchange-correlation potential decays with a_X/r instead of the correct $1/r$ behavior. In most range-separated hybrid functionals, the short-range part is treated by an exchange functional and the long-range part by 100% exact exchange. Some common examples of this type of functionals include ωB97[130], ωB97X[130], CAM-B3LYP[131], LC-ωPBE[132,133], etc.

Functionals belonging to the fifth and final rung of Jacob's ladder use the virtual Kohn−Sham orbitals when calculating the correlation energy. Various approaches have been proposed in the literature, including methods based on perturbation theory[134,135] and random-phase approximation[136]. The most commonly used approach is the double-hybrid density functionals[137,138] (DHDFs). In DHDF, a part of the correlation energy is computed using the MP2 correlation energy expression computed from the KS orbitals:

$$E_{XC}^{DHDF} = (1 - a_X)E_X^{(meta-)GGA} + a_X E_X^{HF} + (1 - a_C)E_C^{(meta-)GGA} + a_C E_C^{MP2} \tag{2.72}$$

One example of a popular DHDF is the B2PYLP[139] functional, which employs 53% exact exchange (or $a_X = 0.53$) and 27% MP2 correlation (or $a_C = 0.27$).

Despite their success, DFT methods suffer from several problems due to the approximate nature of exchange-correlation functionals.[140] For instance, most GGA type functionals severely underestimate reaction barriers and are unsuitable for modeling transition states and chemical reactions. A shortcoming that is relevant to this dissertation is that most DFAs are highly unreliable when it comes to the calculation of non-covalent interaction energies.[141] The interaction energy between two molecules should decrease as the sixth power of the distance in the long-distance limit.[142] Most GGA and hybrid functionals do not capture this long-range dependence resulting in difficulties especially for systems with dominant dispersion interactions. The simplest way to overcome this problem is to add an energy correction E^{disp} to the base DFT energy E^{DFT}:

$$E_{total} = E^{DFT} + E^{disp} \tag{2.73}$$

A semi-empirical way to obtain E^{disp} is via atom-pairwise dispersion energy correction of Grimme[143], which will be discussed in Section 2.3.1. Another popular way to evaluate the E^{disp} is via the exchange-hole dipole moment (XDM) dispersion model developed by Johnson and Becke.[144] Other approaches to correct for incomplete treatment of dispersion interactions also exist and are described in detail in recent review articles[141,145–148].

Most DFAs also suffer from self-interaction error[149–152], which arises from the spurious interaction of an electron with itself. This error comes from the fact that DFT does not treat individual electrons in the same way as wavefunction theory. For a one-electron system like H_2^+, there is zero electron correlation. Therefore, in terms of the energy functionals shown in Equations 2.63 and 2.65, the correlation term $E_C[\rho(r)]$ should be zero and the classical electron-electron repulsion term $J[\rho(r)]$ should cancel the exchange term $E_X[\rho(r)]$ exactly. Most DFT based methods fail to reproduce this behavior. The effects of self-interaction error can also observed in multi-electron systems.[153–155] For multi-electron systems, self-interaction error manifests as an error that has been known in the literature as delocalization error[156–158]. Delocalization error causes DFAs to overstabilize fractional charges and predict spuriously low energies in systems where the charge distribution is delocalized. Delocalization error impacts the calculation of some properties in closed-shell organic molecules.[159–163] A prototypical example of delocalization error can be seen using [10]-annulene as a model.[164,165] It has been shown that [10]-annulene has a handful of conformers out of which specifically one is planar and has extensive electron delocalization and one is non-planar and has no electron delocalization. Accurate wavefunction theory methods predict the planar conformer to be about 6 kcal/mol higher in energy than the non-planar conformer. However, GGA and meta-GGA functionals predict that the planar conformer is lower in energy than the non-planar conformer by about 8 kcal/mol, thereby resulting in a total error of about 14 kcal/mol.

Many of the problems associated with DFT functionals can be understood in terms of delocalization error. However, one another area where DFT functionals appear to fail is the case of strongly correlated systems, characterized by the presence of degeneracy or near degeneracy, having large static correlation.[166] The error associated with strongly correlated systems has been known in the literature as the static correlation error.[167–174] Such strongly correlated systems in wavefunction theory ansatz require more than one Slater determinant to describe the ground state configuration and pose a challenge for not only single configuration wavefunction methods but also for many DFT functionals. One simple example is the dissociation of the H_2 molecule where delocalization error cannot explain the failures of DFT functionals when the two atoms are stretched towards infinite separation.[168–170] Because of static correlation error observed in many DFT functionals, molecules like He_2^+ and $(H_2O)_2^+$ also have erroneous energies when they are stretched to the dissociation limit.[140,175] Zhang *et al.* very recently also investigated the effects of static correlation error on dissociation energies using several DFT functionals and other prototypical molecules like F_2, HF, and NaF.[174] Some other specific examples of static correlation error includes the problems of DFT functionals in the description of the band structure of materials like Mott insulators[176] and superconducting cuprates[177].

2.2 Computational Chemistry Techniques

2.2.1 Geometry Optimization

Geometry optimization[178,179] is an important step in computational studies of molecular structure and reactivity. In a geometry optimization, a set of initial atomic coordinates for a given molecular system are changed iteratively until a three-dimensional atomic arrangement with minimal energy is found. Before discussing the geometry optimization procedure, let's introduce the concept of potential energy surface[180,181] (PES). The PES is the function that gives the molecular energy in terms of the atomic positions. In the Born−Oppenheimer approximation, the value of the PES at a given molecular geometry can be obtained by solving the corresponding electronic Schrödinger equation. Because translations and whole-body rotations of the molecule do not affect its electronic energy, a PES has $3N - 6$ dimensions if the molecule is non-linear, or $3N - 5$ dimensions if it is linear, where N is the number of atoms.

In molecular modeling applications, it is particularly interesting to find the points in the PES that are minima or first-order saddle points. A minimum on the PES corresponds to an equilibrium geometry of a molecule, which is, at least in principle, energetically stable state. Depending on the size of the molecule, there may be many minima on the PES corresponding to its various stable conformers or isomers. The equilibrium geometry with the lowest energy is the "global energy minimum", while other minima are referred to as the "local energy minima". It is also interesting to know how a system changes between one minimum energy structure to another. When studying chemical reactions, reactants and products correspond to different energy minima on the PES containing both species. A "reaction path" is a path on the PES that connects the reactant and product minima. The highest point along the minimum-energy reaction path between reactant(s) and product(s) is the "transition structure" (TS). The TS is a first-order saddle point (a col) on the PES. According to transition state theory[182,183], the difference in energy between TS and reactants is the energy barrier for the reaction and governs the reaction rate, according to the Eyring equation. The general form of the Eyring equation is written as:

$$k = \frac{k_B T}{h} \exp\left(-\frac{\Delta G^{\ddagger}}{RT}\right) \tag{2.74}$$

where k is the reaction rate constant, k_B is the Boltzmann constant, T is the temperature, h is the Planck's constant, R is the gas constant, and $\Delta G^{\ddagger}$ is the Gibb's free energy of activation. Thus, in order to use a calculated PES to study molecular structure and reactivity, one must locate minima, TS, and reaction paths on it. For this purpose, it is necessary to perform geometry optimizations to determine the minima and saddle points on the PES.

The atomic forces experienced by the nuclei in a molecule are equal to the negative of the energy gradient with respect to the atomic coordinates. The matrix elements of the energy gradient matrix $\boldsymbol{g}$ is given as:

$$g_i = \left(\frac{\partial E}{\partial \vec{r}_i}\right) \tag{2.75}$$

Since the atomic forces (or the energy gradient) are zero at minima and TS, they are referred to as the "stationary points" on the PES. To distinguish between a minimum and a saddle point, the matrix of second derivatives of the energy with respect to the atomic coordinates, referred to as the Hessian matrix $\boldsymbol{H}$, is calculated. The equation for the Hessian matrix elements is:

$$H_{ij} = \left(\frac{\partial^2 E}{\partial \vec{r}_i \partial \vec{r}_j}\right) \tag{2.76}$$

If the Hessian matrix is transformed to mass-weighted coordinates and diagonalized, then the eigenvectors are the normal modes of vibration, and the eigenvalues are proportional to the squares of the vibrational frequencies, provided that the eigenvalues with respect to translation and/or rotation of the molecule are set to zero (see Section 2.2.2). The nature of the stationary point found is a minimum if all the eigenvalues are positive (or frequencies are real numbers). Whereas, it is a TS if only one of the eigenvalues is negative (or the corresponding frequency is an imaginary number due to the square-rooting of a negative eigenvalue) and rest all are positive (or frequencies are real numbers). The eigenvector for the negative eigenvalue points along the reaction path.

The geometry optimization procedure starts by choosing a coordinate system; several options are available. An initial geometry is also required. The iterative procedure of geometry optimization consists of: (i) calculating the energy and the necessary energy derivatives, (ii) modification of the geometry to step towards the nearest stationary point, and (iii) checking for convergence. Therefore, the ability of a geometry optimization to converge to a sensible stationary point depends on the quality of the starting geometry and algorithmic details such as how the search direction and step size are determined, as well as the convergence settings.

Geometry optimization algorithms vary depending on the information that is available during the optimization procedure (see references 184–187 and references therein). For example, the simplex and univariate search methods are classified as energy-only methods where the only information required by the optimization algorithm is the system's energy. However, such methods require many iterations to achieve satisfactory convergence. The convergence rate can be significantly improved by using gradient based algorithms, i.e., by using the gradient at the current geometry as well as the energy. The steepest descent and conjugate gradient algorithms are examples of gradient based methods. A third-class of algorithms are Hessian based methods like Newton−Raphson and Gill−Murray methods. The Hessian based methods require calculating the Hessian matrix at each geometry in addition to the energy and gradient. Hessian based methods are not as common as gradient based method because calculating the Hessian is computationally expensive. However, these methods are helpful when convergence difficulties are encountered using the other algorithms.

One of the popular optimization methods used in computational chemistry is the quasi-Newton method. The quasi-Newton method is based on the Newton−Raphson method, where the Hessian matrix, instead of being calculated at each iteration, is approximated at the beginning by some inexpensive QM method and then updated regularly. The quasi-Newton method is comparable in computational cost to gradient based methods and comparable in convergence rate to Hessian based methods. In the plain Newton−Raphson method, the Taylor series expansion of the energy of PES around the current molecular geometry $\boldsymbol{x}_i$ is truncated after the first three terms to yield:

$$E(\boldsymbol{x}) = E(\boldsymbol{x}_i) + \boldsymbol{g}_i^T(\boldsymbol{x} - \boldsymbol{x}_i) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_i)^T \boldsymbol{H}_i(\boldsymbol{x} - \boldsymbol{x}_i) \tag{2.77}$$

where, $\boldsymbol{x}$ is a vector with the $3N$ Cartesian coordinates (where N is the number of atoms), $E(\boldsymbol{x})$ is the energy, $\boldsymbol{g}_i^T$ is the $3N$ element gradient vector at $\boldsymbol{x}_i$, and $\boldsymbol{H}_i$ is the $3N \times 3N$ Hessian matrix also at $\boldsymbol{x}_i$. A quasi-Newton optimization begins with an approximate Hessian $\boldsymbol{B}$ (to be distinguished from the actual Hessian $\boldsymbol{H}$). It uses the same quadratic approximation as in Newton−Raphson but using the approximate Hessian to estimate the position of the stationary point:

$$E(\boldsymbol{x}) = E(\boldsymbol{x}_i) + \boldsymbol{g}_i^T(\boldsymbol{x} - \boldsymbol{x}_i) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_i)^T \boldsymbol{B}_i (\boldsymbol{x} - \boldsymbol{x}_i) \tag{2.78}$$

$$\boldsymbol{g}_{i+1} = \boldsymbol{g}(\boldsymbol{x}_{i+1}) = \boldsymbol{g}_i + \boldsymbol{B}_i(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i) = \boldsymbol{0} \tag{2.79}$$

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \alpha \boldsymbol{B}_i{}^{-1} \boldsymbol{g_i} \tag{2.80}$$

The $-\alpha \boldsymbol{B}_i{}^{-1}\boldsymbol{g_i}$ from Equation 2.80 is the step, which determines the direction and magnitude by which the molecular geometry is modified. At any optimization step, the approximate Hessian $\boldsymbol{B}$ or its inverse $\boldsymbol{B}^{-1}$ must satisfy the so-called quasi-Newton condition for the current cycle, i.e., $\boldsymbol{B}_i{}^{-1}(\boldsymbol{g_i} - \boldsymbol{g_{i-1}}) = \boldsymbol{x}_i - \boldsymbol{x}_{i-1}$ and sometimes the so-called heredity property for the previous cycle, i.e., $\boldsymbol{B}_i{}^{-1}(\boldsymbol{g_k} - \boldsymbol{g_{k-1}}) = \boldsymbol{x}_k - \boldsymbol{x}_{k-1}, k < i$. Depending on the desired stationary point (minima or TS), the $\boldsymbol{B}$ or $\boldsymbol{B}^{-1}$ in the quasi-Newton method is updated by any of the various methods known as Symmetric rank one (SR1) or Murtagh−Sargent (MS) update[188], Davidon−Fletcher−Powell (DFP) update[189], Broyden−Fletcher−Goldfarb−Shanno (BFGS) update[190–193], Schlegel update[194], Powell−symmetric−Broyden (PSB) update[195], Bofill update[196,197], Frakas−Schlegel update[198], etc. In all cases, the Hessian update requires only gradient and energy information.

Because the quadratic approximation of the PES (Equation 2.77) in the quasi-Newton method is crude, several optimization cycles are required to reach a stationary point. The use of a Newton−Raphson step when the PES is significantly non-quadratic can lead to large changes to the molecular geometry in the wrong direction. This instability of the quasi-Newton method can be mitigated by various techniques such as the trust radius method (TRM[179,185]), rational function optimization (RFO[179,199–203]), etc. The number of iterations required to reach convergence can also be reduced by using a scheme where a new geometry is constructred as a linear combination of previous geometries so as to minimize (in a least-squares sense) the size of the step (i.e., the residual error in the iterative solution of Equation 2.80), e.g., in geometry optimization by direct inversion in the iterative subspace (GDIIS[204,205]) and geometry optimization by energy-represented direct inversion in the iterative subspace (GEDIIS[206]). A hybrid approach where different optimization methods are used at various stages of convergence can also lead to significant computational savings, particularly for large molecular systems.[206]

2.2.2 Molecular Vibrations

Molecular vibrations[207] are collective motions of atoms in molecule subject to the condition that the center of mass of the molecule remains unchanged and there are no whole-body rotations. Molecular vibrations are described in terms of normal modes that are determined as eigenvectors of a matrix related to the Hessian matrix $\boldsymbol{H}$ (Equation 2.76) at the equilibrium geometry, as described below. Normal modes involve the cooperative movement of a set of atoms in the molecule and, depending on the type of motion, they receive names such as stretching, bending, rocking, wagging, twisting, out-of-plane, etc. In the harmonic approximation, the vibration along different normal modes associated are independent of each other. The assumption is that the normal modes follow Hooke's Law and that the PES is a quadratic function (parabola) in the vicinity of a stationary point.

A non-linear molecule with N atoms has in total $3N - 6$ vibrational normal modes since six degrees of freedom are used to account for translational and rotational motion. Generally, a geometry optimization procedure is followed by a calculation of the vibrational normal modes, which requires calculating the Hessian matrix, in order to characterize the nature of a stationary point. The energy (electronic plus nuclear repulsion) calculated at a stationary point corresponds to a hypothetical and motionless state at 0K. In reality, molecules at a finite temperature have additional translational, rotational, and vibrational motions that contribute to their free energy. Molecular vibration frequencies are required for the calculation of the thermal contributions to thermodynamic quantities, which are obtained by using statistical thermodynamics[208]. This allows comparing calculated energies to experimental data. Additionally, the vibrational frequencies associated with the normal modes can also be compared with the results of some spectroscopic experiments.

The vibrational normal modes and the associated vibrational frequencies are computed from a set of molecular coordinates and obtained using the Hessian matrix $\boldsymbol{H}$ (Equation 2.76), calculated at the end of the optimization procedure. In practice, to obtain the normal modes and frequencies, the mass-weighted Hessian matrix $\boldsymbol{F}$ is computed. The elements of the mass-weighted Hessian matrix $\boldsymbol{F}$ are:

$$F_{ij} = \frac{1}{\sqrt{m_i m_j}} H_{ij} \tag{2.81}$$

The matrix $\boldsymbol{F}$ is then diagonalized. The eigenvectors are the normal modes of vibration and the eigenvalues (λ_i) are proportional to the squares of the vibrational frequencies. If the Hessian is defined in terms of the $3N$ atomic coordinates, then six eigenvalues are zero and are discarded as they correspond to the translational and rotational motions. The vibrational frequencies ν_i are obtained from the eigenvalues using:

$$\nu_i = \frac{\sqrt{\lambda_i}}{2\pi} \tag{2.82}$$

As mentioned above, an equilibrium geometry (energy minimum) has positive eigenvalues for all normal modes, and therefore all the vibrational frequencies are real numbers. A first-order saddle point or TS must have a negative eigenvalue for one of the normal modes and positive eigenvalues for all remaining normal modes. A TS has one imaginary vibrational frequency, and all other vibrational frequencies are real. The normal mode corresponding to the imaginary vibrational frequency is the transition vector and points along the reaction path connecting the reactant(s) and product(s).

2.2.3 Gas-phase Molecular Properties

A feature of computational chemistry is that it can be used to calculate various thermodynamic properties. This is done by using the principles of statistical mechanics, which act as a link between quantum mechanics and classical thermodynamics and can be used to predict the behavior of macroscopic properties from the statistical behavior of ensembles of microscopic entities.[208] In the following we present a summary of the equations used for computing thermochemical data. Within statistical thermodynamics, the fundamental starting point is the canonical partition function Q, from which all thermodynamic properties of a closed system can be calculated. The canonical partition function is a function of the number of moles, the system volume, and the temperature, $Q(N, V, T)$ and for a system of non-interacting particles it can be written as:

$$Q = \sum_{i}^{states} e^{-\beta E_i} \tag{2.83}$$

where β is a constant and E_i is the total energy of the system in state i. For a system of non-interacting particles which is also indistinguishable, the canonical partition function is:

$$Q = \frac{q^N}{N!} \tag{2.84}$$

where q is the molecular partition function ($q = \sum_i^{mol.\ states} e^{-\beta\epsilon_i}$) and N is the number of molecules in the system. Equation 2.84 is valid for most systems at relatively high temperatures so that all molecules can occupy different states.

Because the different molecular modes of motion are approximately separable, the molecular partition function q can be obtained from the product of individual partition functions for the electronic, translational, rotational, and vibrational motion:

$$q = q_{elec} q_{trans} q_{rot} q_{vib} \tag{2.85}$$

The equation describing the electronic partition function q_{elec} is given as follows:

$$q_{elec} = \sum_{i=0} \omega_i e^{-\varepsilon_i / k_B T} \tag{2.86}$$

where ω_i is the degeneracy of the i-th energy level with energy ε_i, k_B is Boltzmann's constant, and T is the temperature. For most molecules, the electronic energy levels are separated enough that only the ground state contributes to the partition function. If we set the ground state energy $\varepsilon_0 = 0$, Equation 2.86 simplifies to:

$$q_{elec} = \omega_0 \tag{2.87}$$

The translational partition function q_{trans} is:

$$q_{trans} = \left(\frac{2\pi m k_B T}{h^2}\right)^{3/2} V \tag{2.88}$$

where m is the mass of the molecule, h is the Planck's constant, and V is the volume.

The form of the rotational partition function q_{rot} depends on the geometry of a molecule. For a single atom $q_{rot} = 1$. For a linear molecule, the rotational partition function is:

$$q_{rot} = \frac{1}{\sigma_r}\left(\frac{T}{\Theta_r}\right) \tag{2.89}$$

where σ_r is the symmetry number for rotation, equal to the number of rotations that leave the molecule unchanged. Θ_r is the rotational temperature, equal to $h^2/8\pi^2 I k_B$, where I is the moment of inertia. For a non-linear molecule, the rotational partition function is given as:

$$q_{rot} = \frac{\sqrt{\pi}}{\sigma_r}\left(\frac{T^{3/2}}{\sqrt{\Theta_{r,x}\Theta_{r,y}\Theta_{r,z}}}\right) \tag{2.90}$$

where $\Theta_{r,x}$, $\Theta_{r,y}$, and $\Theta_{r,z}$ are the rotational temperatures for the corresponding three moments of inertia of the molecule.

Finally, vibrational partition function q_{vib}, calculated in the harmonic approximation, is composed of a product of contributions from each normal mode. There are $3N-6$ (or $3N-5$ for linear molecules) normal modes. Each normal mode has a characteristic vibrational temperature, $\Theta_{v,K} = h\nu_K/k_B$. If the zero of energy is chosen to the bottom of the internuclear potential energy well, then q_{vib} is given as:

$$q_{vib} = \prod_K \frac{e^{-\Theta_{v,K}/2T}}{1-e^{-\Theta_{v,K}/T}} \tag{2.91}$$

Using Equations 2.85–2.91, all thermodynamic quantities can be calculated. The entropy of a system with N moles is calculated using the following relation:

$$S = S_{elec} + S_{trans} + S_{rot} + S_{vib} \tag{2.92}$$

or,

$$S = Nk_B + Nk_B \ln\left(\frac{q}{N}\right) + Nk_B T\left(\frac{\partial \ln q}{\partial T}\right)_V \tag{2.93}$$

Similarly, the internal energy of a molecule is calculated using the following relation:

$$U = U_{elec} + U_{trans} + U_{rot} + U_{vib} \tag{2.94}$$

or,

$$U = Nk_B T^2\left(\frac{\partial \ln q}{\partial T}\right)_V \tag{2.95}$$

The enthalpy and Gibbs free energy are obtained from the following relations:

$$H = U + pV \tag{2.96}$$

$$G = H - TS \tag{2.97}$$

Given that the evaluation of the thermodynamic properties is moderately computationally costly, most of the molecular properties calculated in this dissertation are obtained directly from the electronic energies of a QM calculation. The main properties considered are:

1) Binding energy:
$$E_{complex}^{bind} = E_{complex} - \sum E_{monomer} \tag{2.98}$$

where $E_{complex}$ is the electronic energy of a non-covalently bound complex and $E_{monomer}$ present inside the sum represents the electronic energies of the individual monomers in the complex.

2) Conformational energy:
$$E_{A}^{conf} = E_{A} - E_{min} \tag{2.99}$$

where E_A is the electronic energy of the conformer A and E_{min} is the electronic energy of the lowest-energy conformer of the same molecule.

3) Deformation energy:
$$E_{A}^{deform} = E_{A} - E_{eqm} \tag{2.100}$$

where E_A is the electronic energy of molecule A that is deformed along a particular normal mode and E_{eqm} is the electronic energy of the equilibrium structure of the same molecule without any deformation.

4) Isomerization energy:
$$E_{A}^{isom} = E_{A} - E_{min} \tag{2.101}$$

where E_A is the electronic energy of the isomer A and E_{min} is the electronic energy of the lowest-energy isomer of the same molecule.

5) Barrier height energy:
$$E_{rxn}^{barrier} = E_{TS} - \sum E_{R/P} \tag{2.102}$$

where E_{TS} is the electronic energy of the transition structure and $E_{R/P}$ present inside the sum is the electronic energy of the reactant(s) or product(s) depending on whether the calculated barrier is for the forward reaction or the reverse reaction.

6) Reaction energy:
$$E_{rxn}^{reaction} = \sum E_{P} - \sum E_{R} \tag{2.103}$$

where E_P represents the electronic energy of each product molecule and E_R represents the electronic energy of each reactant molecule.

7) Bond separation energy:

$$E_{AB}^{bond} = E_{AB} - E_{A^{\cdot}} - E_{B^{\cdot}} \qquad (2.104)$$

where E_{AB} is the electronic energy of the parent molecule AB and $E_{A^{\cdot}}$, $E_{B^{\cdot}}$ are the electronic energies of the two radical fragments $A^{\cdot}$ and $B^{\cdot}$ generated from the homolytic cleavage of a particular bond in AB. In the literature, Equation 2.104 representing the strength of a chemical bond has also been written in form of analogous equations for bond dissociation enthalpy (H_{AB}^{bond}) or the bond dissociation free energy (G_{AB}^{bond}).

2.3 Practical Approaches

2.3.1 The D3 London Dispersion Correction

Grimme and co-workers developed the D3 correction to incorporate the missing dispersion interactions in HF and DFT methods.[209,210] The D3 correction is a variant of two earlier works[211,212] . In the D3 correction the dispersion energy is calculated from the molecular geometry and added to the base DFT method. In D3 (paired with Becke–Johnson or BJ damping[213–215]), the energy correction term $E^{D3(BJ)}$ is:

$$E_{total}^{D3-corrected} = E^{HF/DFT} + E^{D3(BJ)} \qquad (2.105)$$

$$E^{D3(BJ)} = -\frac{1}{2}\sum_{A\neq B}^{atoms}\left(s_6\frac{C_6^{AB}}{R_{AB}^6 + (a_1R_{AB}^0 + a_2)^6} + s_8\frac{C_8^{AB}}{R_{AB}^8 + (a_1R_{AB}^0 + a_2)^8}\right) \qquad (2.106)$$

where C_6^{AB} and C_8^{AB} refers to the sixth-and eight-order dispersion coefficients for atom pairs AB, R_{AB} is the interatomic distance, R_{AB}^0 is equal to $(C_8^{AB}/C_6^{AB})^{1/2}$, s_6 and s_8 are empirically determined scaling factors, and a_1 and a_2 are fitted parameters. A major benefit of the D3 correction scheme is that it depends only on the molecular geometry and the four pre-determined parameters (i.e., s_6, s_8, a_1, a_2) and therefore adds little computational cost to the calculation. The C_6^{AB} are pre-computed and readily available to be looked up for different coordination numbers around each atom and for various atom pairs.[143] The values for C_8^{AB} can then be obtained from C_6^{AB} using a recursive relation. Empirical parameters of D3 are currently available for many elements of the periodic table and for both HF and DFT methods. It is also readily available for usage in most computational chemistry packages.

2.3.2 The HF-3c Method

The HF-3c method was developed by Sure *et al.* in 2013.[216] It is a low-cost composite approach designed to correct the deficiencies of the underlying small-basis-set HF method by using three atom pairwise correction terms. One of the three correction terms is intended to fix missing dispersion interactions from HF and the remaining two mitigate basis set incompleteness errors due to the small basis set.

The HF-3c method uses a basis set called MINIX, which is similar to the scaled MINI basis set (MINIS) by Huzinaga *et al.*[217] for first three rows of the periodic table. MINIX is equivalent to MINIS for elements $H{-}He$ and $B{-}Ne$. MINIX is also similar to MINIS for other elements between $Li{-}Ar$ but MINIX includes one additional p-type basis function for Li, Be, Na, Mg and one additional d-type basis function for $Al{-}Ar$.

The three correction terms present in HF-3c are the so-called geometrical counterpoise (gCP[218]) and short-ranged basis (SRB[216]) incompleteness corrections along with a refitted D3(BJ) dispersion correction. The refitted parameters for D3(BJ)[209,210,213–215] in the HF-3c are: $s_6 = 1.0$, $s_8 = 0.8777$, $a_1 = 0.4171$, and $a_2 = 2.9149$ Å. The gCP term in HF-3c is a geometrical counterpoise correction for the basis set superposition error adopted from earlier work[218]. The formula for the gCP energy correction is:

$$E^{gCP} = \sigma \sum_{A}^{atoms} \sum_{A \neq B}^{atoms} E_A^{miss} \frac{\exp(-\alpha(R_{AB})^{\beta})}{\sqrt{S_{AB} N_B^{virt}}} \tag{2.107}$$

where σ is a global scaling factor, E_A^{miss} refers to the pre-computed atomic energy difference between a nearly complete basis set and the target MINIX basis set, $\exp(-\alpha(R_{AB})^{\beta})$ acts as a decay function that depends on the interatomic distance R_{AB} between atom pairs AB and the fitted parameters α and β, and the $(S_{AB} N_B^{virt})^{-1/2}$ term is a normalization constant that depends on Slater-type overlap integrals S_{AB} and the number of virtual orbitals N_B^{virt} on atom B. The S_{AB} integrals also depend on a fitted parameter η. The gCP parameters (σ, α, β, η) were obtained via a least-squares fit against counterpoise correction data obtained by the scheme of Boys and Bernadi as described in the original publication[218].

The SRB term is a short-range basis correction term designed to deal with the basis set incompleteness of the MINIX basis set. The formula for the SRB energy correction term is:

$$E^{SRB} = -s \sum_{A}^{atoms} \sum_{A \neq B}^{atoms} (Z_A Z_B)^{3/2} \exp(-\gamma \left(R_{AB}^{0,D3}\right)^{3/4} R_{AB}) \tag{2.108}$$

where, Z_A and Z_B are nuclear charges associated with the atoms A and B, s and γ are fitted parameters with values 0.03 and 0.7, R_{AB} is the interatomic distance, and $R_{AB}^{0,D3}$ are the cutoff radii for the D3 dispersion correction. The s and γ parameters were obtained by fitting against high-level atomic forces in a set of 107 equilibrium structures of small organic molecules.

In total, HF-3c consists of nine empirical parameters, three for the D3(BJ) correction, four in the gCP scheme, and two for the SRB correction. Each of the corrections can be calculated from the geometry alone, thereby incurring negligible computational overhead. HF-3c is available for many elements of the periodic table (hydrogen to xenon) and is implemented in the *ORCA*[219] software package. When the three correction terms present in HF-3c are added to correct the underlying energy of the HF method in the MINIX basis set, the total corrected energy is:

$$E_{total}^{HF-3c} = E^{HF/MINIX} + E^{D3(BJ)} + E^{gCP} + E^{SRB} \tag{2.109}$$

2.4 Computational resources

The research conducted in this dissertation was enabled by the computational resource support provided by WestGrid, Compute Canada, and the University of British Columbia’s advanced research computing services. The various high-performance computing platforms where all calculations were executed include “orcinus” (earlier WestGrid and now University of British Columbia), “cedar” (Westgrid and Compute Canada), and “sockeye” (University of British Columbia). The individual calculations were submitted to the queue system of the computing platforms as jobs with resource requirements of a single CPU computing node with number of processors varying between 1 to 40 CPU cores and allocated random access memory varying between 3 gigabytes to 170 gigabytes. More details associated with the computational resources are also presented in the individual research chapters.

References

(1) Atkins, P. W.; Friedman, R. *Molecular Quantum Mechanics, 5th Edition*; Oxford University Press, 2011.

(2) Levine, I. N. *Quantum Chemistry, 7th Edition*; Pearson, 2014.

(3) Griffiths, D. J.; Schroeter, D. F. *Introduction to Quantum Mechanics, 3rd Edition*; Cambridge University Press, 2018.

(4) Engel, T. *Physical Chemistry: Quantum Chemistry and Spectroscopy, 4th Edition*; Pearson, 2019.

(5) Hehre, W. J.; Radom, L.; Schleyer, P. V. R.; Pople, J. *Ab Initio Molecular Orbital Theory*; John Wiley & Sons Inc., 1986.

(6) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models, 2nd Edition*; John Wiley & Sons Inc., 2004.

(7) Jensen, F. *Introduction to Computational Chemistry, 3rd Edition*; John Wiley & Sons Inc., 2017.

(8) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry : Introduction to Advanced Electronic Structure Theory*; Dover Publications Inc., New York, 1996.

(9) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic-Structure Theory*; John Wiley & Sons Inc., 2000.

(10) Leach, A. *Molecular Modelling: Principles and Applications, 2nd Edition*; Pearson, 2001.

(11) Schrödinger, E. Quantisierung Als Eigenwertproblem. *Ann. Phys.* **1926**, *384* (4), 361–376.

(12) Dirac, P. A. M. *The Principles of Quantum Mechanics, 4th Edition*; Oxford University Press, 1982.

(13) Dirac, P. A. M. A New Notation for Quantum Mechanics. *Math. Proc. Cambridge Philos. Soc.* **1939**, *35* (3), 416–418.

(14) Born, M.; Oppenheimer, R. Zur Quantentheorie Der Molekeln. *Ann. Phys.* **1927**, *389* (20), 457–484.

(15) Hartree, D. R. *The Calculation of Atomic Structures*; John Wiley & Sons Inc., 1957.

(16) Heisenberg, W. Mehrkörperproblem Und Resonanz in Der Quantenmechanik. *Zeitschrift für Phys.* **1926**, *38* (6–7), 411–426.

(17) Dirac, P. A. M. On the Theory of Quantum Mechanics. *Proc. R. Soc. London. Ser. A Math. Phys. Sci.*, **1926**, *112* (762), 661–677.

(18) Slater, J. C. The Theory of Complex Spectra. *Phys. Rev.* **1929**, *34* (10), 1293–1322.

(19) Massimi, M. *Pauli's Exclusion Principle*; Cambridge University Press, 2005.

(20) Hartree, D. R. The Wave Mechanics of an Atom with a Non-Coulomb Central Field Part I Theory and Methods. *Math. Proc. Cambridge Philos. Soc.* **1928**, *24* (1), 89–110.

(21) Hartree, D. R. The Wave Mechanics of an Atom with a Non-Coulomb Central Field Part II Some Results and Discussion. *Math. Proc. Cambridge Philos. Soc.* **1928**, *24* (1), 111–132.

(22) Fock, V. Näherungsmethode Zur Lösung Des Quantenmechanischen Mehrkörperproblems. *Zeitschrift für Phys.* **1930**, *61* (1–2), 126–148.

(23) Kalman, D. Leveling with Lagrange: An Alternate View of Constrained Optimization. *Math. Mag.* **2009**, *82* (3), 186–196.

(24) Roothaan, C. C. J. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.* **1951**, *23* (2), 69–89.

(25) Hall, G. G. The Molecular Orbital Theory of Chemical Valency VIII. A Method of Calculating Ionization Potentials. *Proc. R. Soc. London. Ser. A. Math. Phys. Sci.* **1951**, *205* (1083), 541–552.

(26) Nagy, B.; Jensen, F. Basis Sets in Quantum Chemistry; In *Reviews in Computational Chemistry*, John Wiley & Sons Inc., 2017; pp 93–149.

(27) Jensen, F. Atomic Orbital Basis Sets. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3* (3), 273–295.

(28) *Basis Sets in Computational Chemistry*; Perlt, E., Ed.; Lecture Notes in Chemistry; Springer International Publishing: Cham, 2021; Vol. 107.

(29) Boys, S. F. Electronic Wave Functions - I. A General Method of Calculation for the Stationary States of Any Molecular System. *Proc. R. Soc. London. Ser. A. Math. Phys. Sci.* **1950**, *200* (1063), 542–554.

(30) Taketa, H.; Huzinaga, S.; O-ohata, K. Gaussian-Expansion Methods for Molecular Integrals. *J. Phys. Soc. Japan* **1966**, *21* (11), 2313–2324.

(31) Besalú, E.; Carbó-Dorca, R. The General Gaussian Product Theorem. *J. Math. Chem.* **2011**, *49* (8), 1769–1784.

(32) Hellmann, H.; Kassatotschkin, W. Metallic Binding According to the Combined Approximation Procedure. *J. Chem. Phys.* **1936**, *4* (5), 324–325.

(33) Phillips, J. C.; Kleinman, L. New Method for Calculating Wave Functions in Crystals and Molecules. *Phys. Rev.* **1959**, *116* (2), 287–294.

(34) Hellmann, H. A New Approximation Method in the Problem of Many Electrons. *J. Chem. Phys* **1935**, *3*, 61.

(35) Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. Basis Set Exchange: A Community Database for Computational Sciences. *J. Chem. Inf. Model.* **2007**, *47* (3), 1045–1052.

(36) Pritchard, B. P.; Altarawy, D.; Didier, B.; Gibson, T. D.; Windus, T. L. New Basis Set Exchange: An Open, Up-to-Date Resource for the Molecular Sciences Community. *J. Chem. Inf. Model.* **2019**, *59* (11), 4814–4820.

(37) Hehre, W. J.; Stewart, R. F.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals. *J. Chem. Phys.* **1969**, *51* (6), 2657–2664.

(38) Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1971**, *54* (2), 720–723.

(39) Hehre, W. J.; Ditchfield, K.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XII. Further Extensions of

Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56* (5), 2257–2261.

(40) Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90* (2), 1007–1023.

(41) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, *96* (9), 6796–6806.

(42) Woon, D. E.; Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. IV. Calculation of Static Electrical Response Properties. *J. Chem. Phys.* **1994**, *100* (4), 2975–2988.

(43) Schäfer, A.; Horn, H.; Ahlrichs, R. Fully Optimized Contracted Gaussian Basis Sets for Atoms Li to Kr. *J. Chem. Phys.* **1992**, *97* (4), 2571–2577.

(44) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully Optimized Contracted Gaussian Basis Sets of Triple Zeta Valence Quality for Atoms Li to Kr. *J. Chem. Phys.* **1994**, *100* (8), 5829–5835.

(45) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7* (18), 3297–3305.

(46) Furche, F.; Ahlrichs, R.; Hättig, C.; Klopper, W.; Sierka, M.; Weigend, F. Turbomole. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4* (2), 91–100.

(47) Jensen, F. Polarization Consistent Basis Sets: Principles. *J. Chem. Phys.* **2001**, *115* (20), 9113–9125.

(48) Jensen, F. Polarization Consistent Basis Sets: II. Estimating the Kohn-Sham Basis Set Limit. *J. Chem. Phys.* **2002**, *116* (17), 7372–7379.

(49) Jensen, F. Polarization Consistent Basis Sets. III. The Importance of Diffuse Functions. *J. Chem. Phys.* **2002**, *117* (20), 9234–9240.

(50) Jensen, F. Polarization Consistent Basis Sets. IV. The Basis Set Convergence of Equilibrium Geometries, Harmonic Vibrational Frequencies, and Intensities. *J. Chem. Phys.* **2003**, *118* (6), 2459–2463.

(51) Jensen, F. Segmented Contracted Basis Sets Optimized for Nuclear Magnetic Shielding. *J. Chem. Theory Comput.* **2015**, *11* (1), 132–138.

(52) Pyykkö, P. Relativistic Effects in Structural Chemistry. *Chem. Rev.* **1988**, *88* (3), 563–594.

(53) Pyykkö, P. The Physics behind Chemistry and the Periodic Table. *Chem. Rev.* **2012**, *112* (1), 371–384.

(54) Pitzer, K. S. Relativistic Effects on Chemical Properties. *Acc. Chem. Res.* **1979**, *12* (8), 271–276.

(55) Wang, S. G.; Liu, W.; Schwarz, W. H. E. On Relativity, Bonding, and Valence Electron Distribution. *J. Phys. Chem. A* **2002**, *106* (5), 795–803.

(56) Schwerdtfeger, P.; Pašteka, L. F.; Punnett, A.; Bowman, P. O. Relativistic and Quantum Electrodynamic Effects in Superheavy Elements. *Nucl. Phys. A* **2014**, *944*, 551–577.

(57) Cao, X.; Dolg, M. Pseudopotentials and Modelpotentials. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (2), 200–210.

(58) Dolg, M.; Cao, X. Relativistic Pseudopotentials: Their Development and Scope of Applications. *Chem. Rev.* **2012**, *112* (1), 403–480.

(59) Dolg, M. Relativistic Effective Core Potentials. In *Handbook of Relativistic Quantum Chemistry*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2017; pp 449–478.

(60) Cundari, T. R.; Benson, M. T.; Lutz, M. L.; Sommerer, S. O. Effective Core Potential Approaches to the Chemistry of the Heavier Elements; In *Reviews in Computational Chemistry*, John Wiley & Sons, Ltd, 2007; pp 145–202.

(61) Pyykkö, P. Relativistic Quantum Chemistry. *Adv. Quantum Chem.* **1978**, *11* (C), 353–409.

(62) Huo, W. M.; Kim, Y. K. Use of Relativistic Effective Core Potentials in the Calculation of Total Electron-Impact Ionization Cross-Sections. *Chem. Phys. Lett.* **2000**, *319* (5–6), 576–586.

(63) Odoh, S. O.; Schreckenbach, G. Performance of Relativistic Effective Core Potentials in DFT Calculations on Actinide Compounds. *J. Phys. Chem. A* **2010**, *114* (4), 1957–1963.

(64) Goddard, W. A. New Foundation for the Use of Pseudopotentials in Metals. *Phys. Rev.* **1968**, *174* (3), 659–662.

(65) Kahn, L. R.; Goddard, W. A. Ab Initio Effective Potentials for Use in Molecular Calculations. *J. Chem. Phys.* **1972**, *56* (6), 2702–2712.

(66) Kahn, L. R.; Baybutt, P.; Truhlar, D. G. Ab Initio Effective Core Potentials: Reduction of All-Electron Molecular Structure Calculations to Calculations Involving Only Valence Electrons. *J. Chem. Phys.* **1976**, *65* (10), 3826–3853.

(67) Shaw, R. A.; Hill, J. G. Prescreening and Efficiency in the Evaluation of Integrals over Ab Initio Effective Core Potentials. *J. Chem. Phys.* **2017**, *147* (7), 74108.

(68) McKenzie, S. C.; Epifanovsky, E.; Barca, G. M. J.; Gilbert, A. T. B.; Gill, P. M. W. Efficient Method for Calculating Effective Core Potential Integrals. *J. Phys. Chem. A* **2018**, *122* (11), 3066–3075.

(69) Szalay, P. G.; Müller, T.; Gidofalvi, G.; Lischka, H.; Shepard, R. Multiconfiguration Self-Consistent Field and Multireference Configuration Interaction Methods and Applications. *Chem. Rev*. **2012**, *112*(1), 108–181.

(70) Roca-Sanjuán, D.; Aquilante, F.; Lindh, R. Multiconfiguration Second-Order Perturbation Theory Approach to Strong Electron Correlation in Chemistry and Photochemistry. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*(4), 585–603.

(71) Marian, C. M.; Heil, A.; Kleinschmidt, M. The DFT/MRCI Method. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2019**, *9*(2), e1394.

(72) Lyakh, D. I.; Musiał, M.; Lotrich, V. F.; Bartlett, R. J. Multireference Nature of Chemistry: The Coupled-Cluster View. *Chem. Rev*. **2012**, *112*(1), 182–243.

(73) Park, J. W.; Al-Saadon, R.; Macleod, M. K.; Shiozaki, T.; Vlaisavljevich, B. Multireference Electron Correlation Methods: Journeys along Potential Energy Surfaces. *Chem. Rev.* **2020**, *120*(13), 5878–5909.

(74) Sherrill, C. D.; Schaefer III, H. F. The Configuration Interaction Method: Advances in Highly Correlated Approaches. In *Advances in Quantum Chemistry*; Academic Press Inc., 1999; Vol. 34, pp 143–269.

(75) Bartlett, R. J.; Musiał, M. Coupled-Cluster Theory in Quantum Chemistry. *Rev. Mod. Phys.* **2007**, *79* (1), 291–352.

(76) Shavitt, I.; Bartlett, R. J. *Many–Body Methods in Chemistry and Physics: MBPT and Coupled-Cluster Theory*; Cambridge University Press, 2009; Vol. 9780521818322.

(77) Bartlett, R. J. Coupled-cluster Theory: An Overview of Recent Developments; In *Modern Electronic Structure Theory Part II*, World Scientific, 1995; pp 1047–1131.

(78) Cremer, D. Møller-Plesset Perturbation Theory: From Small Molecule Methods to Methods for Thousands of Atoms. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*(4), 509–530.

(79) Purvis, G. D.; Bartlett, R. J. A Full Coupled-Cluster Singles and Doubles Model: The Inclusion of Disconnected Triples. *J. Chem. Phys.* **1982**, *76* (4), 1910–1918.

(80) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. Quadratic Configuration Interaction. A General Technique for Determining Electron Correlation Energies. *J. Chem. Phys.* **1987**, *87* (10), 5968–5975.

(81) Helgaker, T.; Jørgensen, P.; Olsen, J. Calibration of the Electronic-Structure Models. In *Molecular Electronic-Structure Theory*; John Wiley & Sons, Ltd: Chichester, UK, 2014; pp 817–883.

(82) Leininger, M. L.; Allen, W. D.; Schaefer III, H. F.; Sherrill, C. D. Is Moller - Plesset Perturbation Theory a Convergent Ab Initio Method? *J. Chem. Phys.* **2000**, *112* (21), 9213–9222.

(83) Truhlar, D. G. Basis-Set Extrapolation. *Chem. Phys. Lett.* **1998**, *294* (1–3), 45–48.

(84) Řezáč, J.; Hobza, P. Describing Noncovalent Interactions beyond the Common Approximations: How Accurate Is the "Gold Standard," CCSD(T) at the Complete Basis Set Limit? *J. Chem. Theory Comput.* **2013**, *9* (5), 2151–2155.

(85) Feller, D. Application of Systematic Sequences of Wave Functions to the Water Dimer. *J. Chem. Phys.* **1992**, *96* (8), 6104–6114.

(86) Feller, D. The Use of Systematic Sequences of Wave Functions for Estimating the Complete Basis Set, Full Configuration Interaction Limit in Water. *J. Chem. Phys.* **1993**, *98* (9), 7059–7071.

(87) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. Basis-Set Convergence of Correlated Calculations on Water. *J. Chem. Phys.* **1997**, *106* (23), 9639–9646.

(88) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. Basis-Set Convergence in Correlated Calculations on Ne, N_2, and H_2O. *Chem. Phys. Lett.* **1998**, *286* (3–4), 243–252.

(89) Ten-no, S.; Noga, J. Explicitly Correlated Electronic Structure Theory from R12/F12 Ansätze. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2* (1), 114–125.

(90) Kong, L.; Bischoff, F. A.; Valeev, E. F. Explicitly Correlated R12/F12 Methods for Electronic Structure. *Chem. Rev.* **2012**, *112*(1), 75–107.

(91) Grüneis, A.; Hirata, S.; Ohnishi, Y. ya; Ten-No, S. Perspective: Explicitly Correlated Electronic Structure Theory for Complex Systems. *J. Chem. Phys.* **2017**, *146*, 80901.

(92) Werner, H. J.; Köppl, C.; Ma, Q.; Schwilk, M. Explicitly Correlated Local Electron Correlation Methods. In *Fragmentation: Toward Accurate Calculations on Complex Molecular Systems*; John Wiley & Sons Inc., 2017; pp 1–79.

(93) Ma, Q.; Werner, H. J. Explicitly Correlated Local Coupled-Cluster Methods Using Pair Natural Orbitals. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*(6), e1371.

(94) Tew, D. P.; Klopper, W.; Neiss, C.; Hättig, C. Quintuple-ζ Quality Coupled-Cluster Correlation Energies with Triple-ζ

Basis Sets. *Phys. Chem. Chem. Phys.* **2007**, *9* (16), 1921–1930.

(95) Werner, H.-J.; Knizia, G.; Adler, T. B.; Marchetti, O. Benchmark Studies for Explicitly Correlated Perturbation- and Coupled Cluster Theories. In *Progress in Physical Chemistry Volume 3*; Oldenbourg Wissenschaftsverlag GmbH, 2010; pp 203–221.

(96) Saebø, S.; Pulay, P. Local Treatment of Electron Correlation. *Annu. Rev. Phys. Chem.* **1993**, *44* (1), 213–236.

(97) Riplinger, C.; Neese, F. An Efficient and near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method. *J. Chem. Phys.* **2013**, *138* (3), 034106.

(98) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural Triple Excitations in Local Coupled Cluster Calculations with Pair Natural Orbitals. *J. Chem. Phys.* **2013**, *139* (13), 134101.

(99) Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. Communication: An Improved Linear Scaling Perturbative Triples Correction for the Domain Based Local Pair-Natural Orbital Based Singles and Doubles Coupled Cluster Method [DLPNO-CCSD(T)]. *J. Chem. Phys.* **2018**, *148* (1), 164105.

(100) Karton, A. A Computational Chemist’s Guide to Accurate Thermochemistry for Organic Molecules. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2016**, *6* (3), 292–310.

(101) Koch, W.; Holthausen, M. C. *A Chemist’s Guide to Density Functional Theory*; John Wiley & Sons Inc., 2001.

(102) Sholl, D. S.; Steckel, J. A. *Density Functional Theory: A Practical Introduction*; John Wiley and Sons Inc., 2009.

(103) Burke, K.; Wagner, L. O. DFT in a Nutshell. *Int. J. Quant.Chem.* **2013**, *113*(2), 96–101.

(104) Becke, A. D. Perspective: Fifty Years of Density-Functional Theory in Chemical Physics. *J. Chem. Phys.* **2014**, *140* (18), 18A301.

(105) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136* (3B), B864.

(106) Thomas, L. H. The Calculation of Atomic Fields. *Math. Proc. Cambridge Philos. Soc.* **1927**, *23* (5), 542–548.

(107) Fermi, E. Eine Statistische Methode Zur Bestimmung Einiger Eigenschaften Des Atoms Und Ihre Anwendung Auf Die Theorie Des Periodischen Systems Der Elemente. *Zeitschrift für Phys.* **1928**, *48* (1–2), 73–79.

(108) Dirac, P. A. M. Note on Exchange Phenomena in the Thomas Atom. *Math. Proc. Cambridge Philos. Soc.* **1930**, *26* (3), 376–385.

(109) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140* (4A), A1133.

(110) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. Prescription for the Design and Selection of Density Functional Approximations: More Constraint Satisfaction with Fewer Fits. *J. Chem. Phys.* **2005**, *123* (6), 6158.

(111) Perdew, J. P. Jacob’s Ladder of Density Functional Approximations for the Exchange-Correlation Energy. In *AIP Conference Proceedings*; AIP Publishing, 2003; Vol. 577, pp 1–20.

(112) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A Look at the Density Functional Theory Zoo with the Advanced GMTKN55 Database for General Main Group Thermochemistry, Kinetics and Noncovalent Interactions. *Phys. Chem. Chem. Phys.* **2017**, *19* (48), 32184–32215.

(113) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115* (19), 2315–2372.

(114) Slater, J. C. A Simplification of the Hartree-Fock Method. *Phys. Rev.* **1951**, *81* (3), 385–390.

(115) Bloch, F. Bemerkung Zur Elektronentheorie Des Ferromagnetismus Und Der Elektrischen Leitfähigkeit. *Zeitschrift für Phys.* **1929**, *57* (7–8), 545–555.

(116) Ceperley, D. M.; Alder, B. J. Ground State of the Electron Gas by a Stochastic Method. *Phys. Rev. Lett.* **1980**, *45* (7), 566–569.

(117) Vosko, S. H.;Wilk, L.;Nusair, M. . Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Can. J. Phys.* **1980**, *58* (8), 1200–1211.

(118) Perdew, J. P.; Wang, Y. Accurate and Simple Analytic Representation of the Electron-Gas Correlation Energy. *Phys. Rev. B* **1992**, *45* (23), 13244–13249.

(119) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865–3868.

(120) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100.

(121) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37* (2), 785–789.

(122) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta–Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91* (14), 146401.

(123) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Constantin, L. A.; Sun, J. Workhorse Semilocal Density Functional for Condensed Matter Physics and Quantum Chemistry. *Phys. Rev. Lett.* **2009**, *103* (2), 026403.

(124) Becke, A. D. A New Mixing of Hartree-Fock and Local Density-Functional Theories. *J. Chem. Phys.* **1993**, *98* (2), 1372–1377.

(125) Langreth, D. C.; Perdew, J. P. Exchange-Correlation Energy of a Metallic Surface: Wave-Vector Analysis. *Phys. Rev. B* **1977**, *15* (6), 2884–2901.

(126) Yang, W. Generalized Adiabatic Connection in Density Functional Theory. *J. Chem. Phys.* **1998**, *109* (23), 10107–10110.

(127) Becke, A. D. Density-functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652.

(128) Adamo, C.; Barone, V. Toward Reliable Density Functional Methods without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.* **1999**, *110* (13), 6158–6170.

(129) Mardirossian, N.; Head-Gordon, M. How Accurate Are the Minnesota Density Functionals for Noncovalent Interactions, Isomerization Energies, Thermochemistry, and Barrier Heights Involving Molecules Composed of Main-Group Elements? *J. Chem. Theory Comput.* **2016**, *12*(9), 4303–4325.

(130) Chai, J. Da; Head-Gordon, M. Systematic Optimization of Long-Range Corrected Hybrid Density Functionals. *J. Chem. Phys.* **2008**, *128* (8), 084106.

(131) Yanai, T.; Tew, D. P.; Handy, N. C. A New Hybrid Exchange–Correlation Functional Using the Coulomb-Attenuating Method (CAM-B3LYP). *Chem. Phys. Lett.* **2004**, *393* (1–3), 51–57.

(132) Vydrov, O. A.; Scuseria, G. E. Assessment of a Long-Range Corrected Hybrid Functional. *J. Chem. Phys.* **2006**, *125* (23), 234109.

(133) Vydrov, O. A.; Heyd, J.; Krukau, A. V.; Scuseria, G. E. Importance of Short-Range versus Long-Range Hartree-Fock Exchange for the Performance of Hybrid Density Functionals. *J. Chem. Phys.* **2006**, *125* (7), 074106.

(134) Görling, A.; Levy, M. Correlation-Energy Functional and Its High-Density Limit Obtained from a Coupling-Constant Perturbation Expansion. *Phys. Rev. B* **1993**, *47* (20), 13105–13113.

(135) Görling, A.; Levy, M. Exact Kohn-Sham Scheme Based on Perturbation Theory. *Phys. Rev. A* **1994**, *50* (1), 196–204.

(136) Eshuis, H.; Bates, J. E.; Furche, F. Electron Correlation Methods Based on the Random Phase Approximation. *Theor. Chem. Acc.* **2012**, *131* (1), 1–18.

(137) Martin, J. M. L.; Santra, G. Empirical Double-Hybrid Density Functional Theory: A 'Third Way' in Between WFT and DFT. *Israel J. Chem* **2020**, *60*(8-9), 787–804.

(138) Goerigk, L.; Grimme, S. Double-Hybrid Density Functionals. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4* (6), 576–600.

(139) Grimme, S. Semiempirical Hybrid Density Functional with Perturbative Second-Order Correlation. *J. Chem. Phys.* **2006**, *124* (3), 034108.

(140) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112* (1), 289–320.

(141) DiLabio, G. A.; Otero-de-la-Roza, A. Noncovalent Interactions in Density Functional Theory. In *Reviews in Computational Chemistry*; wiley, 2016; Vol. 29, pp 1–97.

(142) Stone, A. *The Theory of Intermolecular Forces*; Oxford University Press, 2013.

(143) Goerigk, L. A Comprehensive Overview of the DFT-D3 London-Dispersion Correction. In *Non-Covalent Interactions in Quantum Chemistry and Physics: Theory and Applications*, Elsevier Inc., 2017, pp 195–219.

(144) Johnson, E. R. The Exchange-Hole Dipole Moment Dispersion Model. In *Non-Covalent Interactions in Quantum Chemistry and Physics: Theory and Applications*; Elsevier Inc., 2017; pp 169–194.

(145) Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C. Dispersion-Corrected Mean-Field Electronic Structure Methods. *Chem. Rev.* **2016**, *116* (9), 5105–5154.

(146) Grimme, S. Density Functional Theory with London Dispersion Corrections. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (2), 211–228.

(147) Hermann, J.; DiStasio, R. A.; Tkatchenko, A. First-Principles Models for van Der Waals Interactions in Molecules and Materials: Concepts, Theory, and Applications. *Chem. Rev.* **2017**, *117*(6), 4714–4758.

(148) Stöhr, M.; Van Voorhis, T.; Tkatchenko, A. Theory and Practice of Modeling van Der Waals Interactions in Electronic-

Structure Calculations. *Chem. Soc. Rev.* **2019**, *48*, 4118–4154.

(149) Zhang, Y.; Yang, W. A Challenge for Density Functionals: Self-Interaction Error Increases for Systems with a Noninteger Number of Electrons. *J. Chem. Phys.* **1998**, *109* (7), 2604–2608.

(150) Merkle, R.; Savin, A.; Preuss, H. Singly Ionized First-Row Dimers and Hydrides Calculated with the Fully-Numerical Density-Functional Program NUMOL. *J. Chem. Phys.* **1992**, *97* (12), 9216–9221.

(151) Perdew, J. P.; Levy, M. Comment on "Significance of the Highest Occupied Kohn-Sham Eigenvalue." *Phys. Rev. B - Condens. Matter Mater. Phys.* **1997**, *56* (24), 16021–16028.

(152) Savin, A. In *Recent Developments and Applications of Modern Density Functional Theory*; Seminario, J. M., Ed.; Elsevier, Amsterdam, 1996; p 327.

(153) Vydrov, O. A.; Scuseria, G. E.; Perdew, J. P. Tests of Functionals for Systems with Fractional Electron Number. *J. Chem. Phys.* **2007**, *126* (15), 154109.

(154) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E. Density Functionals That Are One- and Two- Are Not Always Many-Electron Self-Interaction-Free, as Shown for H2+, He2+, Li H+, and Ne2+. *J. Chem. Phys.* **2007**, *126* (10), 104102.

(155) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E. Spurious Fractional Charge on Dissociated Atoms: Pervasive and Resilient Self-Interaction Error of Common Density Functionals. *J. Chem. Phys.* **2006**, *125* (19), 194112.

(156) Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Localization and Delocalization Errors in Density Functional Theory and Implications for Band-Gap Prediction. *Phys. Rev. Lett.* **2008**, *100* (14), 146401.

(157) Johnson, E. R.; Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Delocalization Errors in Density Functionals and Implications for Main-Group Thermochemistry. *J. Chem. Phys.* **2008**, *129* (20), 204112.

(158) Gani, T. Z. H.; Kulik, H. J. Where Does the Density Localize? Convergent Behavior for Global Hybrids, Range Separation, and DFT+U. *J. Chem. Theory Comput.* **2016**, *12* (12), 5931–5945.

(159) Otero-de-la-Roza, A.; Johnson, E. R.; DiLabio, G. A. Halogen Bonding from Dispersion-Corrected Density-Functional Theory: The Role of Delocalization Error. *J. Chem. Theory Comput.* **2014**, *10* (12), 5436–5447.

(160) Heaton-Burgess, T.; Yang, W. Structural Manifestation of the Delocalization Error of Density Functional Approximations: C4N+2 Rings and C20 Bowl, Cage, and Ring Isomers. *J. Chem. Phys.* **2010**, *132* (23), 234113.

(161) Whittleton, S. R.; Sosa Vazquez, X. A.; Isborn, C. M.; Johnson, E. R. Density-Functional Errors in Ionization Potential with Increasing System Size. *J. Chem. Phys.* **2015**, *142* (18), 184106.

(162) Zheng, X.; Liu, M.; Johnson, E. R.; Contreras-García, J.; Yang, W. Delocalization Error of Density-Functional Approximations: A Distinct Manifestation in Hydrogen Molecular Chains. *J. Chem. Phys.* **2012**, *137* (21), 214106.

(163) Otero-De-La-Roza, A.; Johnson, E. R. Analysis of Density-Functional Errors for Noncovalent Interactions between Charged Molecules. *J. Phys. Chem. A* **2020**, *124* (2), 353–361.

(164) Castro, C.; Karney, W. L.; McShane, C. M.; Pemberton, R. P. [10]Annulene: Bond Shifting and Conformational Mechanisms for Automerization. *J. Org. Chem.* **2006**, *71* (8), 3001–3006.

(165) King, R. A.; Crawford, T. D.; Stanton, J. F.; Schaefer III, H. F. Conformations of [10]Annulene: More Bad News for Density Functional Theory and Second-Order Perturbation Theory. *J. Am. Chem. Soc.* **1999**, *121* (46), 10788–10793.

(166) Savin, A. On Degeneracy, near-Degeneracy and Density Functional Theory. *Theor. Comput. Chem.* **1996**, *4*, 327–357.

(167) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Fractional Spins and Static Correlation Error in Density Functional Theory. *J. Chem. Phys.* **2008**, *129* (12), 121104.

(168) Fuchs, M.; Niquet, Y. M.; Gonze, X.; Burke, K. Describing Static Correlation in Bond Dissociation by Kohn-Sham Density Functional Theory. *J. Chem. Phys.* **2005**, *122* (9), 094116.

(169) Peach, M. J. G.; Teale, A. M.; Tozer, D. J. Modeling the Adiabatic Connection in H2. *J. Chem. Phys.* **2007**, *126* (24), 244104.

(170) Baerends, E. J. Exact Exchange-Correlation Treatment of Dissociated H2 in Density Functional Theory. *Phys. Rev. Lett.* **2001**, *87* (13), 133004.

(171) Becke, A. D. A Real-Space Model of Nondynamical Correlation. *J. Chem. Phys.* **2003**, *119* (6), 2972–2977.

(172) Becke, A. D. Real-Space Post-Hartree-Fock Correlation Models. *J. Chem. Phys.* **2005**, *122* (6), 064101.

(173) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Insights into Current Limitations of Density Functional Theory. *Science* **2008**, *31*(5890), 792–794.

(174) Zhang, D.; Truhlar, D. G. Unmasking Static Correlation Error in Hybrid Kohn-Sham Density Functional Theory. *J. Chem. Theory Comput.* **2020**, *16* (9), 5432–5440.

(175) Vande Vondele, J.; Sprik, M. A Molecular Dynamics Study of the Hydroxyl Radical in Solution Applying Self-Interaction-Corrected Density Functional Methods. *Phys. Chem. Chem. Phys.* **2005**, *7*,1363–1367.

(176) Fazekas, P. *Lecture Notes on Electron Correlation and Magnetism*; Series in Modern Condensed Matter Physics; World Scientific, 1999; Vol. 5.

(177) Perry, J. K. Importance of Static Correlation in the Band Structure of High-Temperature Superconductors. *J. Phys. Chem. A* **2000**, *104* (11), 2438–2444.

(178) Schlegel, H. B. Geometry Optimization on Potential Energy Surfaces; In *Modern Electronic Structure Theory Part I*, World Scientific, 1995; pp 459–500.

(179) Schlegel, H. B. Geometry Optimization. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*(5), 790–809.

(180) Heidrich, D.; Kliesch, W.; Quapp, W. *Properties of Chemically Interesting Potential Energy Surfaces*; Lecture Notes in Chemistry; Springer Berlin Heidelberg: Berlin, Heidelberg, 1991; Vol. 56.

(181) Hratchian, H. P.; Schlegel, H. B. Finding Minima, Transition States, and Following Reaction Pathways on Ab Initio Potential Energy Surfaces. In *Theory and Applications of Computational Chemistry*; Elsevier, 2005; pp 195–249.

(182) Laidler, K. J.; King, M. C. The Development of Transition-State Theory. *J. Phys. Chem.* **1983**, *87* (15), 2657–2664.

(183) Anslyn, E. V.; Dougherty, D. A. *Modern Physical Organic Chemistry*; University Science Books, 2006.

(184) Gill, P. E.; Murray, W.; Wright, M. H. *Practical Optimization*; Society for Industrial and Applied Mathematics, 2019.

(185) Fletcher, R. *Practical Methods of Optimization, 2nd Edition*; John Wiley and Sons Inc., 2000.

(186) Dennis, J. E.; Schnabel, R. B. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*; Society for Industrial and Applied Mathematics, 1996.

(187) Scales, L. E. *Introduction to Non-Linear Optimization*; Macmillan Education UK, 1985.

(188) Murtagh, B. A.; Sargent, R. W. H. Computational Experience with Quadratically Convergent Minimization Methods. *Comput. J.* **1970**, *13* (2), 185–194.

(189) Fletcher, R.; Powell, M. J. D. A Rapidly Convergent Descent Method for Minimization. *Comput. J.* **1963**, *6* (2), 163–168.

(190) Broyden, C. G. The Convergence of a Class of Double-Rank Minimization Algorithms 1. General Considerations. *IMA J. Appl. Math.* **1970**, *6* (1), 76–90.

(191) Fletcher, R. New Approach to Variable Metric Algorithms. *Comput. J.* **1970**, *13* (3), 317–322.

(192) Goldfarb, D. A Family of Variable-Metric Methods Derived by Variational Means. *Math. Comput.* **1970**, *24* (109), 23.

(193) Shanno, D. F. Conditioning of Quasi-Newton Methods for Function Minimization. *Math. Comput.* **1970**, *24* (111), 647.

(194) Schlegel, H. B. Optimization of Equilibrium Geometries and Transition Structures. *J. Comput. Chem.* **1982**, *3* (2), 214–218.

(195) M. J. D. Powell, A New Algorithm for Unconstrained Optimation; In *Nonlinear Programming*, J. B. Rosen, O. L. Mangasarian and K. Ritter, Eds. Academic Press, New York, 1970.

(196) Bofill, J. M. Updated Hessian Matrix and the Restricted Step Method for Locating Transition Structures. *J. Comput. Chem.* **1994**, *15* (1), 1–11.

(197) Bofill, J. M. Remarks on the Updated Hessian Matrix Methods. *Int. J. Quantum Chem.* **2003**, *94* (6), 324–332.

(198) Farkas, Ö.; Schlegel, H. B. Methods for Optimizing Large Molecules. II. Quadratic Search. *J. Chem. Phys.* **1999**, *111* (24), 10806–10814.

(199) Simons, J.; Jørgensen, P.; Taylor, H.; Ozment, J. Walking on Potential Energy Surfaces. *J. Phys. Chem.* **1983**, *87* (15), 2745–2753.

(200) Simons, J.; Nichols, J. Strategies for Walking on Potential Energy Surfaces Using Local Quadratic Approximations. *Int. J. Quantum Chem.* **1990**, *38* (24 S), 263–276.

(201) Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. Search for Stationary Points on Surfaces. *J. Phys. Chem.* **1985**, *89* (1), 52–57.

(202) Baker, J. An Algorithm for the Location of Transition States. *J. Comput. Chem.* **1986**, *7* (4), 385–395.

(203) Baker, J. An Algorithm for Geometry Optimization without Analytical Gradients. *J. Comput. Chem.* **1987**, *8* (5), 563–574.

(204) Császár, P.; Pulay, P. Geometry Optimization by Direct Inversion in the Iterative Subspace. *J. Mol. Struct.* **1984**, *114* (C), 31–34.

(205) Farkas, Ö.; Farkas, Ö.; Farkas, Ö.; Schlegel, H. B. Methods for Optimizing Large Molecules. Part III. An Improved Algorithm for Geometry Optimization Using Direct Inversion in the Iterative Subspace (GDIIS). *Phys. Chem. Chem. Phys.* **2002**, *4* (1), 11–15.

(206) Li, X.; Frisch, M. J. Energy-Represented Direct Inversion in the Iterative Subspace within a Hybrid Geometry Optimization Method. *J. Chem. Theory Comput.* **2006**, *2* (3), 835–839.

(207) Cross, P. C.; Wilson Jr., E. B., Decius, J. C. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*; Dover Publications Inc., New York, 1955.

(208) McQuarrie, D. A.; Simon, J. D. *Molecular Thermodynamics*; University Science Books, 1999.

(209) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1456–1465.

(210) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104.

(211) Grimme, S. Accurate Description of van Der Waals Complexes by Density Functional Theory Including Empirical Corrections. *J. Comput. Chem.* **2004**, *25* (12), 1463–1473.

(212) Grimme, S. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27* (15), 1787–1799.

(213) Becke, A. D.; Johnson, E. R. Exchange-Hole Dipole Moment and the Dispersion Interaction. *J. Chem. Phys.* **2005**, *122* (15), 154104.

(214) Becke, A. D.; Johnson, E. R. A Density-Functional Model of the Dispersion Interaction. *J. Chem. Phys.* **2005**, *123* (15), 154101.

(215) Johnson, E. R.; Becke, A. D. A Post-Hartree-Fock Model of Intermolecular Interactions: Inclusion of Higher-Order Corrections. *J. Chem. Phys.* **2006**, *124* (17), 174104.

(216) Sure, R.; Grimme, S. Corrected Small Basis Set Hartree-Fock Method for Large Systems. *J. Comput. Chem.* **2013**, *34* (19), 1672–1685.

(217) Tatewaki, H.; Huzinaga, S. A Systematic Preparation of New Contracted Gaussian-Type Orbital Sets. III. Second-Row Atoms from Li through Ne. *J. Comput. Chem.* **1980**, *1* (3), 205–228.

(218) Kruse, H.; Grimme, S. A Geometrical Correction for the Inter- and Intra-Molecular Basis Set Superposition Error in Hartree-Fock and Density Functional Theory Calculations for Large Systems. *J. Chem. Phys.* **2012**, *136* (15), 154101.

(219) Neese, F. Software Update: The ORCA Program System, Version 4.0. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8* (1), e1327.

Part II

In this part, a proof-of-concept study for validating the central hypothesis of this dissertation was conducted. It is hypothesized that atom-centered potentials (ACPs) can be used to mitigate the performance shortcomings of a selected low-level of theory and produce results that are closer to those obtained with a higher level of theory but with orders of magnitude less computer time. If the hypothesis is proven to be true, the stage would be set for developing computationally inexpensive quantum mechanical methods.

To test the hypothesis, ACPs were developed for four elements (H, C, N, O) and the molecular reference data used for parameterizing the ACPs were limited to a few thousand non-covalent properties calculated at a high-level of theory. The low-level of theory for which the ACPs were developed is the minimal basis set Hartree–Fock (HF) approach.

Another important research problem explored in this work was finding an efficient and better procedure for ACP parametrization to overcome some of the serious limitations faced in previous ACP development work. For this purpose, a regularized least-squares regression technique was implemented and tested. The new approach is able to handle hundreds of thousands of reference data points. The validation of the technique would benefit later ACP development projects, such as those outlined in Parts IV and V.

The primary data that support the findings of Chapter 3 is provided in Appendix 1 of this dissertation. Other supporting files have also been deposited to the figshare repository and are openly available at the following URL/DOI: https://doi.org/10.6084/m9.figshare.16912201. The reference of the published paper is as follows: Prasad, V. K.; Otero-de-la-Roza, A.; DiLabio, G. A. *J. Chem. Theory Comput.* 2018, 14 (2), 726–738. © Copyright 2018 American Chemical Society. (DOI: 10.1021/acs.jctc.7b01158).

Chapter 3

Atom-centered potentials with dispersion-corrected minimal basis set Hartree-Fock: An efficient and accurate computational approach for large molecular systems

Abstract

We present a computational methodology based on atom-centered potentials (ACPs) for the efficient and accurate structural modeling of large molecular systems. ACPs are atom-centered one-electron potentials that have the same functional form as effective-core potentials. In recent works, we showed that ACPs can be used to produce a correction to the ground-state wavefunction and electronic energy to alleviate shortcomings in the underlying model chemistry. In this work, we present ACPs for H, C, N, and O that are specifically designed to produce accurate non-covalent binding energies and inter- and intra-molecular geometries when combined with dispersion-corrected Hartree–Fock (HF-D3) and a minimal basis-set (scaled MINI or MINIs). For example, the combined HF-D3/MINIs-ACP method demonstrates excellent performance, with a mean absolute error of 0.36 and 0.28 kcal/mol for the S22x5 and S66x8 benchmark sets, respectively, relative to highly-correlated complete-basis-set data. The application of ACPs results in a significant decrease in error compared to uncorrected HF-D3/MINIs for all benchmark sets examined. In addition, HF-D3/MINIs-ACP, has a cost only slightly higher than a minimal-basis-set HF calculation and can be used with any electronic structure program for molecular quantum chemistry that uses Gaussian basis-sets and effective-core potentials.

1. Introduction

Recent interest in the modeling of large supramolecular systems[1-3] and molecular crystals[4-6] with density-functional theory (DFT) has caused a resurgence of low-cost computational approaches for intermolecular interactions[7–9]. These “cheap methods” offer the option of taking a calculated trade-off between accuracy and cost in the spirit of force-field approaches but preserving the generality of an electronic structure calculation, particularly regarding their ability to model chemical reactions. In fact, a persistent challenge in this area is to find a way to accurately model covalent and non-covalent interactions simultaneously, since both play important roles in determining the structure of supramolecular systems. While good relative accuracy for covalent bond breaking and formation can be obtained with DFT and a modest basis-set (B3LYP/6-31G*, for instance, is widely used to successfully elucidate many organic reaction mechanisms), modeling non-covalent interactions requires either high-level wavefunction theory (CCSD(T) with complete-basis-set extrapolation) or dispersion-corrected DFT methods with a very large basis-set.[10-12] Although recent developments have reduced the asymptotic scaling of these techniques[13-16], computationally inexpensive methods are still in high demand[17-19] particularly if they can be combined with the new reduced-scaling techniques, as is the case with methods based on atom-centered potentials (ACPs)[20-25].

DFT-based methods suffer from additional problems that make them inefficient in practice for large systems. In addition to inaccuracies caused by the numerical integration of the exchange-correlation energy, generalized-gradient-approximation (GGA) functionals predict erroneous charge-transfer between molecules or within a single molecule due to delocalization error, which may lead to significantly overestimated binding energies and reaction barriers, and spuriously small band gaps.[26–30] In addition to these inaccuracies, a practical difficulty is that many generalized-gradient-approximation functionals (e.g. BLYP[31–33], PBE[34]) and exchange functionals with a small fraction of exact exchange (e.g. B3LYP[35,36]) have considerable difficulties in arriving at converged solutions of the self-consistent field (SCF) equations for charged systems such as zwitterions, which are essential in the description of proteins. This practical hurdle greatly hinders the application of DFT methods to large supramolecular systems.

Semi-empirical approaches (e.g. AM1[37], PMx[38-40]) based on a minimal-basis-set HF approximation have been used in the simulation of supramolecular systems extensively thanks to their reduced cost. Until recently, these methods were unable to model intermolecular interactions adequately, particularly when dispersion effects are dominant.[9,41-44] The HF-3c method, recently proposed by Sure and Grimme is similar in spirit to traditional semi-empirical methods.[45] Instead of discarding or parametrizing certain two-electron integrals, HF-3c uses minimal-basis-set Hartree-Fock (HF) combined with three ad hoc formulas to account for dispersion (D3-BJ) [46-48], basis-set superposition error (gCP) [49], and short-range covalent over-binding (SRB)[45]: The last two errors are caused by basis-set incompleteness. Although more computationally expensive than a semi-empirical method, HF-3c is substantially cheaper than any electronic structure method that uses more complete-basis-sets, and circumvents the self-consistent field (SCF) convergence problems of GGA and hybrid density functionals in large systems. While its performance for small molecular systems is superior to semi-empirical methods,[45] HF-3c has demonstrated suboptimal performance in the ranking of lattice energies in molecular crystals for molecular crystal structure prediction in the sixth CCDC blind test[50], which indicates that there is still room for improvement.

In previous works, we showed that atom-centered potentials (ACPs) represent a simple and effective means to improve the accuracy of DFT-based methods. By fitting to high-level *ab initio* wavefunction data, ACPs can be developed to correct for missing dispersion physics in conventional DFT functionals (in this context, ACPs were termed dispersion-correcting potentials, DCPs).[20-22] These DCPs also implicitly correct inherent deficiencies in the underlying functional, such as delocalization and basis-set incompleteness error.[23,24] Recently, we proposed a newer and more systematic way to develop ACPs, and showed that they can be used as a computationally inexpensive means of mitigating the effects of extreme basis-set incompleteness. These basis-set incompleteness potentials (BSIPs) allow the use of minimal or small double-zeta basis-sets with conventional DFT functionals to obtain almost complete-basis-set quality molecular properties.[25] Because ACPs have the same form as conventional effective-core potentials (ECPs)[51-53], they can be used in most computational chemistry programs that allow for the use of ECPs without additional changes to the software.

In this work, we develop a fast and accurate HF-based minimal-basis-set method for the calculation of molecular properties in large molecular systems. Grimme's D3 correction with Becke-Johnson (BJ) damping is used to account for the absence of dispersion in HF, and ACPs are used to correct for the remaining errors, the leading contribution to which is severe basis-set incompleteness error from the minimal-basis-set. The method uses the minimal-basis-set of Huzinaga with scaled exponents (MINIs).[54] ACPs are proposed for the H, C, N, and O atoms and their parameters are obtained by fitting to an extensive set of molecular properties determined using either highly-correlated wavefunction methods at the complete-basis-set (CBS) limit or LC-ωPBE-XDM/aug-cc-pVTZ[55] where obtaining reference data using wavefunction theory is not feasible. The HF-D3/MINIs-ACP approach is computationally efficient and shows excellent performance in the modeling of intermolecular and intramolecular energies and geometries, and serves as an initial proof-of-concept result using the H, C, N, and O atoms that indicate that ACPs can be used successfully to develop computationally inexpensive techniques based on minimal-basis-set electronic structure calculations. One negative aspect of ACPs is that for general use they must be developed for each atom of the periodic table. However, in our (unsystematic) experience, HF-D3/MINIs-ACP gives good results provided ACPs are applied to the majority of atoms in the system.

2. Computational Methodology

2.1 Theoretical background

The method employed to develop the ACPs in this work has been described in detail elsewhere[25], and is only briefly reviewed here. ACPs are one-electron potentials with the same mathematical form as effective-core potentials (ECPs)[51-53]:

$$V^{ACP}(\boldsymbol{r}) = \sum_{A} U^{A}_{L_A}(r_A) + \sum_{l=0}^{L_A-1} \sum_{m=-l}^{l} |Alm\rangle\, U^{A}_{l}(r_A)\, \langle Alm| \tag{1}$$

with,

$$U^{A}_{l}(r) = \sum_{k} c^{A}_{lk}\, e^{-\xi^{A}_{lk} r_A^2} \tag{2}$$

where A represents the atoms on which the potentials are centered (H, C, N, and O), r_A is the distance from nucleus, and $|Alm\rangle$ are spherical harmonics centered on atom A. We will refer to the L_A term as the "local" angular-momentum channel. The coefficients c^{A}_{lk} and exponents ξ^{A}_{lk} in Eq. 2 are adjustable parameters that are determined via least-squares fitting, as described below. The sum in Eq. 2 runs over the number of Gaussian terms defined for atom A and angular-momentum channel l. Unlike ECPs, ACPs do not replace any electrons of the atoms to which they are applied.

The energy contribution that arises from the application of each ACP is obtained from:

$$E^{ACP}\,(\boldsymbol{c}, \boldsymbol{\xi}\,) = \sum_{Alk} c^{A}_{lk}\, \Delta E^{A}_{lk}\,(\xi^{A}_{lk}) = \boldsymbol{c} \cdot \boldsymbol{\Delta E}(\boldsymbol{\xi})^{\intercal} \tag{3}$$

where the ACP terms ΔE^{A}_{lk} are independent of the coefficients to first order in the perturbation induced by the ACP,

$$\Delta E^{A}_{lk}\,(\xi^{A}_{lk}) = \sum_{i} \langle \psi_i| \left(\sum_{m=-l}^{l} |Alm\rangle\ \exp(-\xi^{A}_{lk}\ r_A^2\,)\,\langle Alm| \right) |\psi_i\rangle \tag{4}$$

The $\boldsymbol{c}$ and $\boldsymbol{\Delta E}(\boldsymbol{\xi})^{\intercal}$ are the corresponding ACP coefficients and energy term vectors, and ψ_i's are the self-consistent ground-state Kohn–Sham orbitals.

The fact that the first-order perturbation energy term arising from the application of $V^{ACP}(\boldsymbol{r})$ (Eq. 1) is linear in the coefficients is key to our ACP development method.[25] For all molecules in the training set, the ACP terms ($\Delta E^{A}_{lk}\,(\xi^{A}_{lk})$) are computed for a pre-determined set of exponents designed to affect the region of space relevant to correct the desired molecular properties. The coefficients are determined through a least-squares procedure based on a least-absolute-shrinkage-and-selection-operator (LASSO) method by Tibshirani[56] in which we perform variable selection to determine the optimal exponents to use in our ACPs. At the same time, the magnitude of the coefficients are constrained to ensure that the contribution of second- and higher-order perturbation terms are negligible. We use the name "non-linearity error" to describe the combined contribution of the second- and higher-order perturbation terms.[25]

Other important features of the ACPs are: (i) the use of angular projection operators in the potential allows for the introduction of local anisotropic corrections to respond to changes in the chemical environment of a given atom; (ii) the exponential decrease in the ACP as a function of distance from the nucleus ensures that the effect of a given ACP is localized in the vicinity of the atom; and (iii) the cost of the one-electron integrals associated with the ACPs is negligible compared to SCF calculation, provided an efficient implementation is used. All calculations in this work use the Gaussian 09 program[57].

2.2. Training data sets

For the purpose of fitting our ACPs, a training set composed of several benchmark sets from the literature was assembled. Only molecules containing H, C, N, and O atoms are used in the set. The molecular properties targeted by the training set include non-covalent binding energies, conformational energies of amino acid dimers and trimers, and molecular deformation energies in small organic molecules. A detailed list of the subsets in our training set is given in Table 1. In total, the training set comprises 9814 data points, including 3235 non-covalent binding energies (S22x5[58,59], S66x8[43,60-62], ACHC[63,64], BBI[63,65], and SSI[63,65] sets), 1599 conformational energies (DIPEPCONF and P26[66] sets), and 4980 covalent molecular deformation energies (MOLdef set).

Table 1. Subsets of the training set used for the ACP fit.[a]

Data Set	Class	Num.	Description	Ref. level	Ref.
S22x5	NCI	110	Potential energy curves of small non-covalently interacting dimers	CCSD(T)/CBS	58, 59
S66x8	NCI	528	Potential energy curves of small non-covalent dimers	CCSD(T)/CBS	43, 60-62
ACHC	NCI	54	Interaction energies of adenine-cytosine nucleobase stacking configurations	DW-CCSD(T**)-F12/aug-cc-pVDZ	63, 64
BBI[b]	NCI	94	Interaction energies of dipeptide backbone-backbone complexes	DW-CCSD(T)-F12/aug-cc-pV(D+d)z	63, 65
SSI[b,c]	NCI	2449	Interaction energies of amino acid side chain-side chain complexes	DW-CCSD(T)-F12/aug-cc-pV(D+d)z	63, 65
P26[d]	CONF	69	Conformational energies of five isolated small peptides containing aromatic side chains	CCSD(T)/CBS	66
DIPEPCONF[b]	CONF	1530	Relative energies of ten conformers for each of 153 dipeptide combinations	LC-ωPBE-XDM/aug-cc-pVTZ	This work

Data Set	Class	Num.	Description	Ref. level	Ref.
MOLdef	DEF	4980	Molecular deformation energies relative to the equilibrium geometry of 49 small molecules	LC-ωPBE-XDM/aug-cc-pVTZ	This work

[a] The classes are: non-covalent binding energies (NCI), conformational energies (CONF), and DEF (molecular deformation energies). The "Num." column indicates number of data points present in the set, and "Ref. Level" is the calculation level of the reference data. [b] Excluding the dimers containing methionine and cysteine, which contain sulfur. [c] Excluding the dimers with charged monomers. [d] Excluding WGG10 and WGG12 conformers because of a possible error in the data provided in the supporting information of the cited reference.

Most of the subsets in Table 1 are from the literature. For the present work, two new sets were developed called DIPEPCONF and MOLdef. In both sets, LC-ωPBE-XDM/aug-cc-pVTZ was used as the reference method. The rationale for this choice is that given the size of these sets, running wavefunction theory calculations would be too computationally expensive, but LC-ωPBE-XDM/aug-cc-pVTZ is expected to have a much higher accuracy than HF-D3/MINIS or any ACP-corrected version for the systems in these sets.[55] To further justify its use, we checked the performance of LC-ωPBE-XDM/aug-cc-pVTZ on several benchmark sets for conformational energies from the literature. The mean absolute errors (MAEs) are: ACONF[67,68] (alkane conformations), 0.12 kcal/mol; PCONF[63,67,69] (peptide conformations), 0.61 kcal/mol; SCONF[67,70] (sugar conformational energies), 0.24 kcal/mol. PCONF is a subset of the previously proposed P26 set[66], for which LC-ωPBE-XDM/aug-cc-pVTZ has an MAE of 0.52 kcal/mol. For comparison, LC-ωPBE-XDM/aug-cc-pVTZ gives MAEs of 0.27 kcal/mol and 0.18 kcal/mol for the S22[58] and S66[60] sets, and 0.23 kcal/mol and 0.15 kcal/mol for the S22x5[59] and S66x8[43,61,62] sets of non-covalent binding energies, respectively.

DIPEPCONF is a new dataset for dipeptide conformational energies proposed in this work. The dipeptides in the DIPEPCONF set contain only neutral side chains and are capped with acetate (N-terminal, ACE) and primary amide (C-terminal, NHE) groups. The initial geometries of the dipeptides were generated using the *tleap* tool in *Amber16*[71]. Molecular dynamics (MD) simulations for each amino acid dimer were carried out in the gas-phase using the *ff14SB* force field[72,73], with a heating step of 200 picoseconds followed by a production run of 4200 picoseconds, from which structures were extracted at uniform time intervals to generate a total of 4000 conformers. Each conformer was then subjected to energy minimization using the same force field. The 4000 conformers for each dipeptide were separated into energy bins and eleven conformers were selected spanning a range of energies. These conformers were then subjected to a single point calculation at the LC-ωPBE-XDM/aug-cc-pVTZ level of theory in order to generate the conformational energy reference data. In total, the DIPEPCONF dataset contains 10 conformational energies (11 conformations) for each of the 153 dipeptide sequences considered. Future work is in progress to extend this set, and this will be published elsewhere.

The MOLdef set contains reference data for molecular deformation energies. For 49 small molecules containing only H, C, N, and O, the equilibrium geometries and vibrational frequencies and normal modes were determined using LC-ωPBE-XDM/aug-cc-pVTZ. The geometry of these molecules was then deformed along each calculated normal mode. For each deformation, the reference energy is the LC-ωPBE-XDM/aug-cc-pVTZ energy difference between the deformed and the equilibrium structures. The intent behind fitting our ACPs to this set is to correct for the erroneous intramolecular geometries that arise from using HF combined with a minimal-basis-set, particularly the spuriously short covalent bonds. Three deformations on each side of the equilibrium geometry were considered along each normal mode, such that the relative energy of the distorted structure never exceeded a few dozen kcal/mol. In total, the MOLdef set consists of 4980 relative energies. The Cartesian coordinates along with the reference energies can be found in the Supporting Information (SI).

Each subset in the training set is assigned a weight for the ACP least-squares fit in order to account for the variable magnitude of the numerical values and number of data points. The weight of each subset in Table 1 is calculated using the formula:

$$w_i = \frac{1}{M_i \times N_i} \tag{5}$$

where M_i is the mean absolute value of the reference energies and N_i is the number of data points for subset i. The weights for all subsets are normalized, and the weighted root-mean-square (wRMS) is defined as:

$$\text{wRMS} = \sqrt{\frac{\sum_i^{\text{subsets}} \sum_j^{N_i} w_j \left(y_{\text{ref},j} - y_{\text{HF-D3/MINIs-ACP},j}\right)^2}{\sum_i^{\text{subsets}} N_i}} \tag{6}$$

with $y_{\text{ref},j}$ the reference energy and $y_{\text{HF-D3/MINIs-ACP},j}$ the energy given by the HF-D3/MINIs-ACP method. The wRMS is minimized as a function of the ACP coefficients c_{lk}^{A} in our least-squares fitting procedure.

2.3 ACP development

Angular momentum channels for all ACPs are considered up to the maximum angular momentum primitive in the MINIs basis-set, including the local (i.e. local and s for H; and local, s and p for C, N, and O). The ACP exponents are 0.01, from 0.02 to 0.28 in 0.02 steps, from 0.40 to 2.00 in 0.20 steps, and from 2.50 to 5.00 in 0.50 steps. The ACP terms (ΔE_{lk}^{A} (ξ_{lk}^{A})) corresponding to each atom, angular momentum channel, and exponent were determined for each entry in the training set using the Hartree-Fock (HF) method with the MINIs minimal-basis-set and the D3 dispersion correction. The D3 parameters correspond to those for the HF/aug-cc-pVTZ method and with Becke-Johnson damping: $s_6 = 1.0$, $s_8 = 0.9171$, $a_1(\text{BJ}) = 0.3385$, $a_2(\text{BJ}) = 2.8830$ Å. To find the optimal ACP, we need to determine the subset of all 330 calculated ACP terms that gives good performance as measured by the magnitude of the wRMS, and at the same time yields coefficients that are small enough that the non-linearity error is small. In this way, we ensure that the statistics resulting from our least-squares fit are a faithful representation of the actual performance of the ACP, and that non-linearity error will not be a large contribution to the method's performance in actual applications.

In our previous work on ACPs for basis-set incompleteness error (BSIPs, Ref. 25), we used an iterative procedure by which all combinations of ACP terms for each atom were explored in turn. Although this method resulted in ACPs with good performance, it is also time consuming and limited by the maximum number of ACP terms in each atom, since the number of combinations increases factorially with this value. In this work, we used an alternative procedure based on the least-absolute-shrinkage-and-selection-operator (LASSO) method by Tibshirani.[56] In LASSO, the least-squares function (in our case, the wRMS in Eq. 6) is minimized subject to the condition that l_1-norm of the ACP coefficients does not exceed a certain bound chosen beforehand. The l_1-norm is given as,

$$\|c\|_1 = \sum_i |c_i| \tag{7}$$

This allows us to constrain the ACP fits to give coefficients as small as we choose. In addition, LASSO also performs variable selection, i.e., for the given constraint on the l_1-norm of the coefficients, LASSO automatically selects the best subset of ACP terms and assigns zero coefficients to the rest. The ACPs determined using the LASSO method have lower wRMS than using the method in Ref. 25. Perhaps more importantly, the fit takes minutes, instead of days, which enables the use of much larger training sets, and the ACPs are not limited to have a certain number of terms per atom. In this work, we used the local linearization plus active set method proposed by Osborne et al.[74] and implemented in octave/MATLAB by Mark Schmidt[75,76]. After some exploration, we determined that a l_1-norm bound of 5.0 a.u. on the coefficients is a good compromise between accuracy and non-linearity error.[22]

2.4 Validation data sets

In addition to the training set, we use several other benchmark sets from the literature to validate the performance of the developed ACPs. The subsets of this validation set have been selected to test non-covalent binding energies and relative conformational energies, and are shown in Table 2.

Table 2. Subsets of the validation set.[a]

Group	Subset	Num.	Description	Ref. level	Ref.
Non-covalent interaction energies					
HYDROCARBONS	HC12	12	Interaction energies of saturated and unsaturated hydrocarbon dimers	CCSD(T)/CBS	77
	ADIM6	6	Interaction energies of n-alkane dimers	CCSD(T)/CBS	67, 78
	$CH_4 \cdot$ PAH	382[c]	Interaction energies of methane with polycyclic aromatic hydrocarbons	CCSD(T)/CBS	63, 79, 80
	$C_2H_4 \cdot$ NT	75	Interaction energies of ethene with coronene	CCSD(T)/CBS	63
CO_2-CAPTURE	$CO_2 \cdot$ PAH	249	Interaction energies of CO_2 with polycyclic aromatic hydrocarbons	CCSD(T**)-F12avg/CBS	63, 81
	$CO_2 \cdot$ NPHAC	96	Interaction energies of CO_2 with nitrogen-doped poly heterocyclic aromatic compounds	CCSD(T)/CBS	63, 82
LARGE-SYSTEMS	S12L[b]	10	Interaction energies of large host-guest supramolecular motifs	corrected expt.	1, 2, 83-86
	S30L[b]	23			
WATER	SHIELDS38	38	Interaction energies of water clusters, $(H_2O)_n$, with n=2-10	CCSD(T)/CBS	87
CHARGED	IONICHB	120	Dissociation curves of small, charged, hydrogen-bonded complexes	CCSD(T)/CBS	43
	SSI (charged)[b]	766	Dimers of amino acid side chain-side chain complexes having charged monomers only	DW-CCSD(T)-F12/aug-cc-pV(D+d)z	63, 65
BIOMOLECULES	A24[b]	19	Interaction energies of small non-covalently bound complexes	CCSD(T)/CBS	12
	HSG	21	Model protein-ligand interaction energies	CCSD(T)/CBS	88, 89
	HBC6	118	Dissociation curves of doubly hydrogen-bonded complexes	CCSD(T)/CBS	88, 90
Relative conformational energies					
CONFORMERS	ACONF	15	Conformational energies of n-alkanes	W1h-val	67, 68
	BCONF	64	Conformational energies of butane-1,4-diol	CCSD(T)-F12b/cc-pVTZ-F12	91
	MCONF	51	Conformational energies of melatonin	CCSD(T)/CBS	92
	PCONF	10	Conformational energies of Phenyl-Glycyl-Glycine tripeptide	CCSD(T**)-F12a/CBS	63, 67, 69
	SCONF	17	Conformational energies of two carbohydrates	CCSD(T)/CBS	67, 70
	TRCONF	8	Conformational energies of two tetrapeptides	CCSD(T)/CBS	93

[a] the “Num.” column indicates number of data points and “Ref. level” is the calculation level of the reference data.[b] Only molecules containing H, C, N and O atoms. [c] 23 out of 405 geometries were missing from the supporting information of Ref. 63.

3. Results

3.1 Optimized ACPs for HF-D3/MINIs

The ACP exponents and coefficients resulting from the fit are listed in Table 3. In general, ACP terms with higher exponents have higher coefficients in absolute value because the corresponding potential term reaches farther away from the atom, and therefore gives a higher energy contribution. Unlike the BSIPs presented in Ref. 25, the ACPs given in Table 3 were optimized for HF-D3/MINIs against high-level reference data, so they cannot be used with other methods or basis-sets because they correct not only for basis-set-incompleteness but also for deficiencies in HF-D3.

The constraint for the LASSO fit was chosen to give an ACP that minimizes the wRMS in a self-consistent calculation over the training set. The number of terms per atom in this ACP is automatically determined by LASSO, and therefore the optimized ACPs in Table 3 contain many more terms than those we have developed previously. A sample input file demonstrating the use of the ACPs in the Gaussian program is provided in the SI.

Table 3. HF-D3/MINIs atom-centered potentials for the H, C, N, and O atoms.[a]

Atom (A)	**Function type (l)**	ξ_{lj}^{A}	c_{lj}^{A}
H	local	0.01	-0.00001938089
		0.02	-0.00003366414
		0.04	0.001323631399
		0.06	-0.00300563983
		0.10	0.00090072893
		0.12	0.00141564539
		0.22	0.00858410759
		0.40	-0.025198082700
		1.00	0.03040038924
	s	0.01	0.00897184727
		0.02	-0.03953828144
		0.04	0.04867972215
		0.06	0.03449941507
		0.10	-0.01590602911
		0.14	-0.13674945686
		0.40	0.40214230125
		2.50	-1.03353005814
C	local	0.01	0.00001861609
		0.02	0.00021332142
		0.04	-0.00214933468

Atom (A)	Function type (l)	ξ_{lj}^{A}	c_{lj}^{A}
		0.06	0.00516749173
		0.10	-0.01555553001
		0.16	0.05344004796
		0.26	-0.13758669254
		0.60	0.23120921343
		1.40	-0.80389740259
	s	0.01	-0.01552134192
		0.02	0.00045549650
		0.16	0.02692938642
		0.24	0.01943925488
		0.26	0.03167389276
	p	0.02	0.01111334155
		0.08	-0.01209320774
		0.18	0.00852089081
		0.20	0.05995111066
		0.22	0.01316335258
		0.24	0.12664813920
		1.20	-0.23218368640
N	local	0.01	0.00005935026
		0.02	0.00014788866
		0.04	-0.00030466045
		0.06	0.00062390375
		0.10	-0.00803928317
		0.16	0.01653732266
		0.60	-0.14314848707
	s	0.01	0.00262789783
		0.04	-0.03994239966
		0.06	-0.07041918366
		0.08	-0.03080252469
		0.40	0.07565676804
	p	0.01	-0.01376692773
		0.02	0.01612955962
		0.04	0.00621017916
		0.06	0.06263892557
		0.22	0.04296706278
		1.00	-0.01832723217
		1.20	-0.24431696314
O	local	0.01	-0.00016818935
		0.02	0.00060565334
		0.04	-0.00330007999
		0.06	0.01108579450
		0.08	-0.00827068380
		0.10	-0.00526053319
		0.28	0.00491351734
		0.80	-0.27851873745
		1.00	-0.03196194558
	s	0.01	0.02609885310
		0.02	0.00355349168
		0.04	0.03976597732
		0.10	0.07826701426
	p	0.02	-0.02356374593

Atom (*A*)	Function type (*l*)	ξ_{lj}^{A}	c_{lj}^{A}
		0.04	-0.00042034334
		0.14	-0.00601202155
		0.20	0.03760145141
		0.26	0.02986288464
		0.28	0.02034075728

[a] ξ_{lj}^{A} and c_{lj}^{A} indicate the ACP exponents and coefficients, respectively.

The first step in the validation of our ACPs is to apply the resulting HF-D3/MINIs-ACP method to the training set to make sure that our least-squares fit is representative of results that would be obtained from self-consistent field (SCF) calculations in which that ACPs are applied and that non-linearity error is not detrimental to the performance of the ACP.[25] By comparing the statistics from the fit and from a self-consistent HF-D3/MINIs-ACP calculation on the training set, we make sure that the contribution from the second- and higher-order terms in Eq. 4 are not significant. Table 4 compares the mean absolute errors (MAEs) obtained from the least-squares fit and from the validation calculations. For comparison, the results obtained using HF-D3/MINIs, HF-D3/aug-cc-pVTZ (abbreviated aTZ), and with the HF-3c method are also given in the table. The MAEs obtained from the fit deviate by 0.15 kcal/mol or less from the results of the corresponding self-consistent calculations, indicating that the l_1-norm constraint in the LASSO fit was successful in preventing excessive non-linearity error. The MAEs obtained using the uncorrected method range from 0.84 to 3.42 kcal/mol, and application of our ACPs reduce the MAEs by up to a factor of five.

Table 4. Mean absolute errors (MAEs) in kcal/mol with respect to high-level reference data for the various subsets of the training set.[a]

Set	HF-D3/MINIs	ACP-Fit	ACP-SCF	HF-D3/aTZ	HF-3c
S22	2.17	0.28	0.43	1.03	0.53
S22x5	1.40	0.32	0.36	0.70	0.53
S66	1.74	0.21	0.24	0.68	0.38
S66x8	1.24	0.27	0.28	0.51	0.37
ACHC	1.44	0.24	0.27	0.34	0.28
BBI	1.05	0.27	0.22	0.60	0.87
SSI	0.84	0.17	0.17	0.23	0.21
P26	2.02	0.40	0.47	0.68	1.20
DIPEPCONF	2.44	0.81	0.85	0.89	1.15
MOLdef	3.42	0.92	0.90	1.73	2.90

[a] Uncorrected HF-D3/MINIs is compared to the results obtained from the fitting procedure ("ACP-Fit"), and to the MAE from the application of HF-D3/MINIs-ACP in self-consistent field calculations ("ACP-SCF"). For comparison, the MAEs obtained using the HF-3c and HF-D3/aTZ methods are also provided.

Table 4 also compares HF-D3/MINIs-ACP to HF-D3/aTZ, which is close to the complete-basis-set limit, and to the HF-3c method. HF-D3/MINIs-ACP outperforms HF-3c in all subsets of our training set. The improvement relative to HF-3c is quite large in the case of BBI and P26. On the other hand, the performance of the HF-D3/MINIs-ACP (0.28 kcal/mol) and HF-3c (0.37 kcal/mol) methods on the S66x8 set, which was also used as fitting set to obtain the parameters for the gCP correction in the HF-3c approach, is quite similar. It is also interesting to compare the minimal-basis-set methods to HF-D3/aTZ, which is reasonably close to the complete-basis-set limit. The performance of HF-D3/aTZ is rather poor for all subsets, and reflects that the D3 correction provides an energy correction amounting to only about half of the total correlation energy contribution to the properties of the fitting sets. The poor performance of HF-D3/aTZ is particularly evident in the binding energies of small molecular dimers, such as the S22, for which the MAE exceeds 1 kcal/mol (c.f. B3LYP-D3, 0.36 kcal/mol[67]). The MAEs for the S22x5 and S66x8 sets for HF-D3/MINIs-ACP and HF-3c are 0.23/0.34 kcal/mol and 0.14/0.17 kcal/mol, respectively, lower than those of HF-D3/aTZ. These results (indeed all of the results shown in Table 4) indicate that the ACPs and the 3c correction rectify the errors associated with basis-set incompleteness and the partial absence of correlation. In comparison, the MAE in the S22x5 and S66x8 using B3LYP-D3(BJ)/aug-cc-pVTZ is 0.25 kcal/mol and 0.17 kcal/mol, respectively, about 0.1 kcal/mol lower than HF-D3/MINIs-ACP.[94]

Unlike HF-D3/MINIs-ACP, however, HF-3c is significantly worse than HF-D3/aTZ for conformational and, particularly, molecular deformation energies. These results are somewhat surprising because the gCP and SRB were not designed, in principle, for the purpose of rectifying the poor performance of HF-D3 for non-covalent binding energies, but do in fact perform this function. Despite the fact that the SRB was specifically fit to reproduce high-level intramolecular geometries, the MAE for the molecular deformations is quite high. This has a noticeable impact on the calculated molecular frequencies and intramolecular geometries. All molecules in the MOLdef set were relaxed using HF-D3/MINIs, HF-D3/MINIs-ACP, and HF-3c, and the resulting geometries compared to the LC-ωPBE-XDM/aug-cc-pVTZ equilibrium geometries using Kabsch's algorithm[95]. The average root-mean-square deviation (RMSD) of the atomic coordinates are 0.0586 (HF-D3/MINIs), 0.0379 (HF-D3/MINIs-ACP), and 0.0464 (HF-3c). HF-3c and HF-D3/MINIs-ACP both improve HF-D3/MINIs, but the latter gives better geometries. The difference between these three methods is even more striking for the vibrational frequencies: The mean absolute percent errors for the vibrational frequencies in the MOLdef set are 13.4% for uncorrected HF-D3/MINIs, 5.2% for HF-D3/MINIs-ACP, and 14.3% for HF-3c. Therefore, although the SRB term in HF-3c goes a small way towards repairing the intramolecular geometries, the quality of the calculated vibrational frequencies decreases. In contrast, ACPs are very successful in correcting the significant errors in HF-D3/MINIs frequencies.

The signed errors and their averages for the four methods described above are shown in Figure 1, which display strip-charts in which all of the signed error data are plotted. In all cases, the ACPs correction is successful in reducing the spread of the errors relative to both HF-D3/MINIs and HF-D3/aTZ. The performance of HF-3c is also excellent with a small bias in all sets except BBI and MOLdef, and a spread of the error that is in general smaller than uncorrected HF-D3/MINIs. The HF-3c results show a larger spread of errors than HF-D3/MINIs-ACP for conformational energies and molecular deformations.

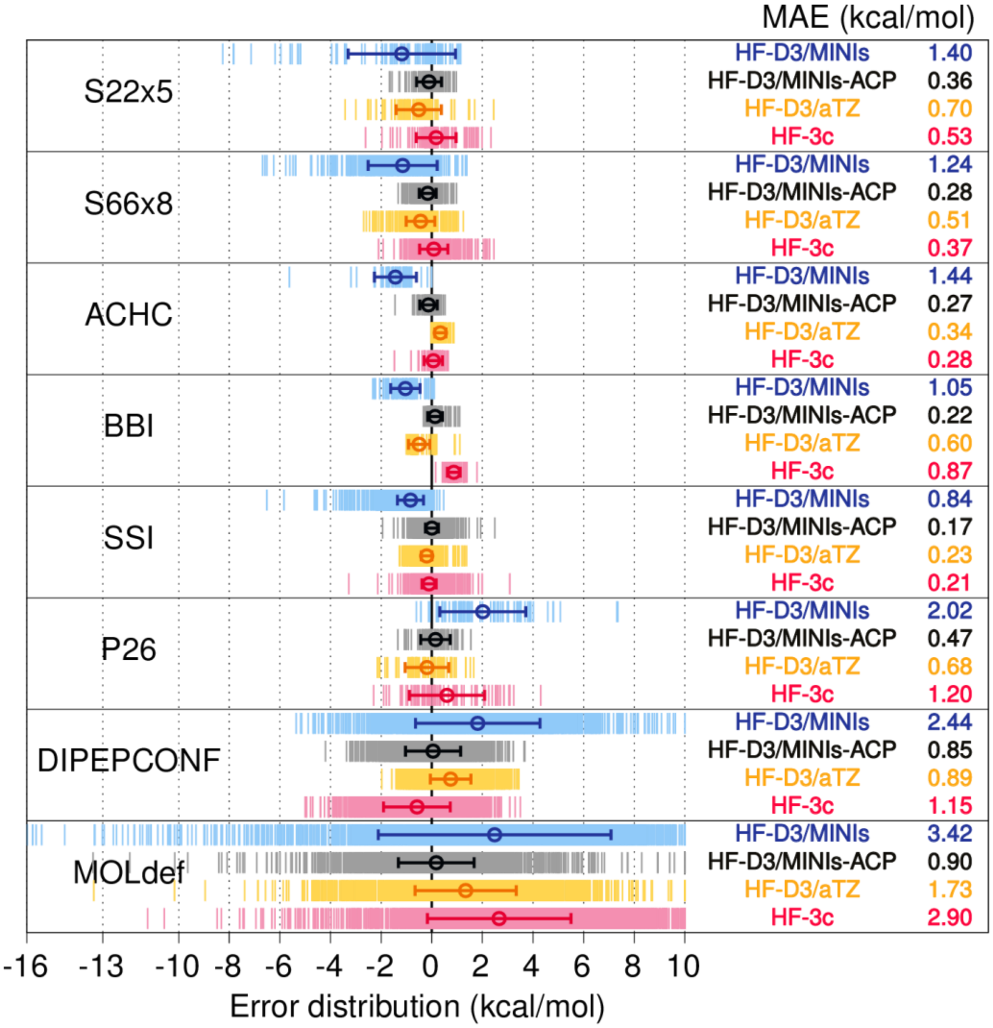


Figure 1. Signed errors associated with the subsets of the training set in Table 1 using HF-D3/MINIs corrected with ACPs (HF-D3/MINIs-ACP) and the 3c correction (HF-3c). The HF-D3/aTZ results are also given. The circle represents the mean error (ME) and the error bar is the standard deviation of the error. The numbers on the right are mean absolute error (MAE) in kcal/mol.

3.2 Performance of HF-D3/MINIs-ACP on the validation set

The performance of HF-D3/MINIs-ACP compared to HF-3c and uncorrected HF-D3/MINIs for the validation set is shown in Table 5. The error distribution, mean errors, and standard deviation for each method and subset are shown in Figure 2.

Table 5. Mean absolute errors (MAEs, in kcal/mol) for the various subsets of the validation set using uncorrected HF-D3/MINIs, and the same method with ACPs (HF-D3/MINIs-ACP) and the 3c correction approach (HF-3c).

Group	Subset	HF-D3/MINIs	HF-D3/MINIs-ACP	HF-3c
HYDROCARBONS		**1.29**	**0.35**	**0.25**
	HC12	1.80	0.26	0.42
	ADIM6	1.90	0.21	0.47
	CH_4 · PAH	1.19	0.29	0.19
	C_2H_4 · NT	1.71	0.67	0.48
CO_2-CAPTURE		**1.72**	**0.88**	**0.57**
	CO_2 · PAH	1.64	0.87	0.55
	CO_2 · NPHAC	1.92	0.89	0.63
LARGE-SYSTEMS		**15.49**	**8.36**	**6.09**
	S12L	15.93	10.27	6.28
	S30L	15.54	7.53	6.01
WATER	SHIELDS38	**30.62**	**4.99**	**7.67**
CHARGED		**3.25**	**1.95**	**2.41**
	IONICHB	4.45	2.47	2.68
	SSI (charged)	3.07	1.87	2.38
BIOMOLECULES		**2.75**	**0.69**	**1.00**
	A24	0.73	0.32	0.44
	HSG	1.69	0.62	0.74
	HBC6	3.27	0.76	1.13
CONFORMERS		**2.27**	**0.69**	**1.05**
	ACONF	1.44	0.98	0.89
	BCONF	2.40	0.50	0.58
	MCONF	0.88	0.71	0.89
	PCONF	2.43	0.50	2.28
	SCONF	5.20	1.17	1.47
	TRCONF	5.14	0.80	3.64

Table 5 and Figure 2 show that, in all cases, the application of the ACPs improves the performance of HF-D3/MINIs. The aggregate MAEs for the binding energies (BEs) and conformational energies are reduced by a factor of 1.9 (LARGE-SYSTEMS) to 6.1 (WATER). This level of performance is similar to

HF-3c (improvement factors of 1.3 - 5.2), however the two methods differ in the kinds of systems whose properties are most accurately predicted.

The HF-D3/MINIs-ACP approach improves the BEs in hydrocarbons group (HYDROCARBONS) on average by a factor of about 4, but the performance for the interactions between ethylene and carbon nanotubes ($C_2H_4 \cdot NT$) and to a lesser extent between methane and polycyclic aromatic hydrocarbons ($CH_4 \cdot PAH$) is comparatively worse than the other two subsets of the HYDROCARBONS group. In these subsets, the error is dominated by the non-equilibrium dimers, particularly those in which the two monomers are at shorter distances than at the equilibrium geometry. The HF-D3/MINIs-ACP approach also shows a more modest improvement of a factor of about 2 over uncorrected HF-D3/MINIs for more specialized sets like CO_2-CAPTURE, which consists of systems relevant in carbon dioxide capture studies. This is encouraging because carbon dioxide model systems were not a part of the training set.

Similar observations can be made about the water clusters in the SHIELDS38 set, composed of $(H_2O)_n$ clusters with n=2-10. The ACPs were fitted only to water dimers, but the statistics show a generalized improvement of binding energies for larger water clusters, with the MAE reduced by a factor of 5 compared to uncorrected HF-D3/MINIs. Figure 2 shows that for this set in particular there is a strong over-binding tendency in HF-D3/MINIs caused by basis-set incompleteness error. This effect is corrected by the application of the ACPs, although the over-binding bias is not completely removed. While the ACPs presented in this work are meant to be general, specialized ACPs can also be fitted for particular systems of interest, in the spirit of force field development techniques. For instance, in a recent article,[96] we showed that ACPs fitted to the SHIELDS38 set and a collection of 2-body, 3-body, and 4-body contributions to the binding energy can correct the shortcomings of the BH&HLYP-XDM/aug-cc-pVTZ method to such a degree that the performance of the ACP-corrected method greatly exceeds wavefunction theory results. This is an interesting feature of ACPs that can be useful when extreme accuracy and computational efficiency for a single system is required, such as in molecular dynamics studies of homogeneous substances or crystal structure prediction.

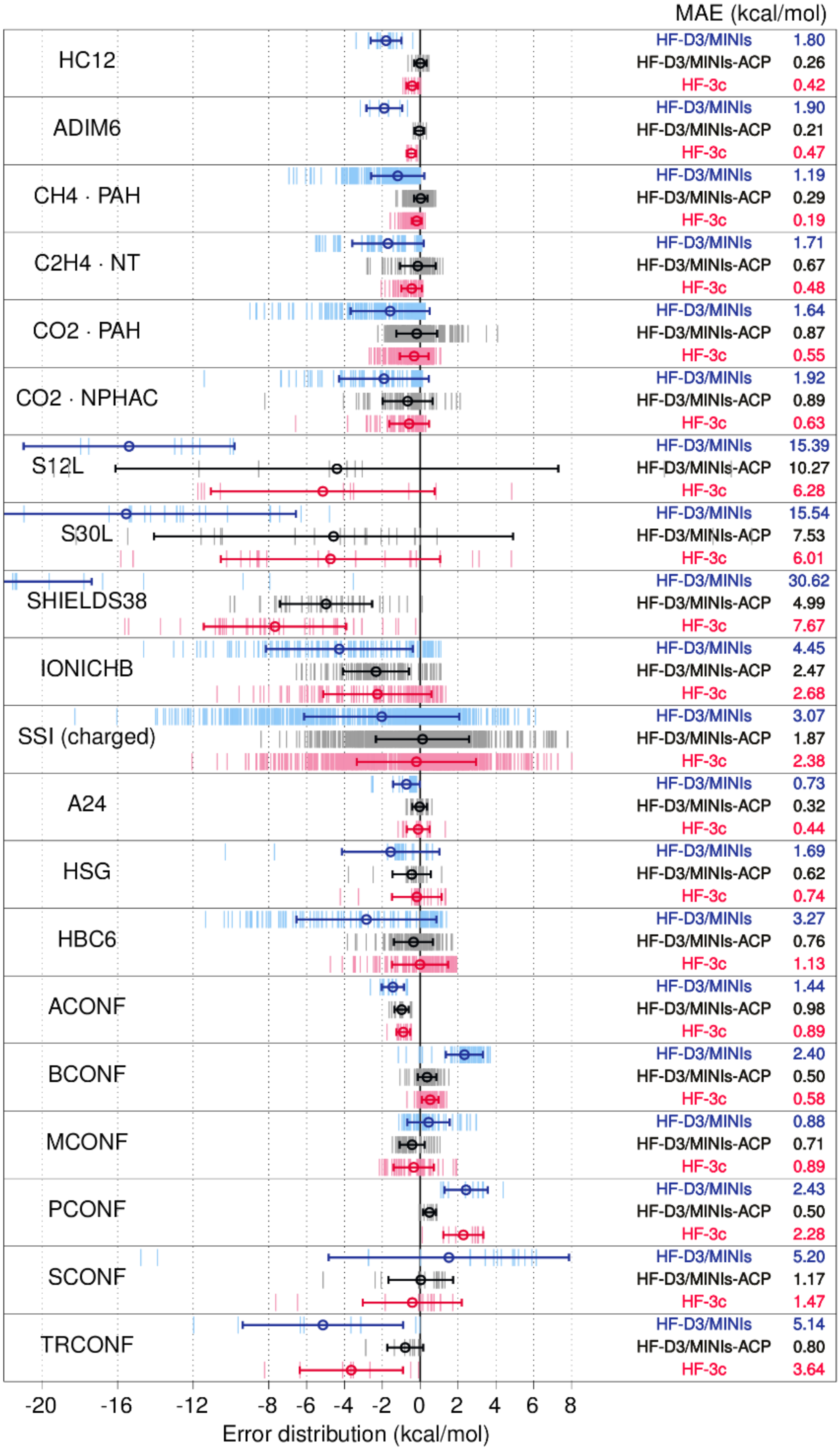
The reduction of the MAE resulting by the application of our ACPs to the large host-guest complexes of S12L and S30L sets (LARGE-SYSTEMS group) and to the charged systems (CHARGED group) is more modest. The S12L, which is contained in the S30L but does not feature any hydrogen-bonded systems, is special in that the performance of various dispersion-corrected functionals show an unusual dependence on the base functional.[3] For instance, while BLYP-XDM, B3LYP-XDM, or LC-ωPBE-XDM routinely outperform PBE-XDM in the description of binding energies in small molecular dimers, it is the latter that performs best for the S12L set, with an MAE of 1.5 kcal/mol[3] (c.f. BLYP, 4.2; B3LYP, 4.0; LC-ωPBE-XDM, 6.8 kcal/mol). The reason for this dependence is unknown at present, but candidates for an explanation are a favorable error cancellation in PBE (hydrogen-bonded systems, for which PBE is notoriously bad, are absent from the S12L) or errors from the methods used in the back-correction of the experimental results from which the reference data were derived. HF-D3/MINIs for the S12L has an MAE of 15.9 kcal/mol, which is reduced to only 10.27 kcal/mol upon application of the ACPs (6.28 in the case of HF-3c). In the S30L, ACPs reduce the MAE from 15.5 kcal/mol to 7.5 kcal/mol (c.f. 6.0 kcal/mol with HF-3c), indicating that they are more successful in representing the hydrogen bonded systems in the S30L not present in the S12L. For comparison, Brandenburg *et al.* reported a MAE of 6.6 kcal/mol using PBE-D3 at an estimated complete-basis-set limit for the S30L.[8]

The MAE for the charged systems is reduced from 3.25 kcal/mol for HF-D3/MINIs to 1.95 kcal/mol upon application of the ACPs (c.f. 2.41 for HF-3c). The charged systems in the SSI were purposefully left out of the training set, since it is clear that a minimal-basis-set does not have enough flexibility to describe anionic systems. It is also important to note that charged systems are present in the S12L, S30L, and HSG, and the errors for those systems are significantly higher than for the rest of the dimers in these sets. For instance, the MAE for HSG using HF-D3/MINIs-ACP drops from 0.62 kcal/mol to 0.29 kcal/mol when the charged systems are removed from the set. For comparison, Burns *et al.* reported a MAE of 0.48 kcal/mol for the whole HSG set using B3LYP-D3/aug-cc-pVTZ[97] and Torres and DiLabio reported an MAE of 0.15 kcal/mol using B3LYP-DCP/6-31+G(2d,2p)[98]. In spite of this, the application of ACPs is still beneficial, even for the subsets composed solely of charged systems—the MAE is decreased by a factor slightly smaller than 2 both in the IONICHB and the charged systems of the SSI set.

The application of HF-D3/MINIs-ACP to the molecular dimers with importance in biological systems (BIOMOLECULES group) improved the BEs on average by a factor of about 4. The MAE for the HBC6 subset of BIOMOLECULES obtained with HF-D3/MINIs-ACP is the largest amongst all of the subsets at 0.76 kcal/mol. Nevertheless, this level of performance is quite good, suggesting that HF-D3/MINIs-ACP approach may offer a faster alternative to accurately model non-covalent interactions in larger sized molecules of biological significance. For comparison, Burns *et al.* reported MAEs for HBC6 of 0.55 and 1.12 kcal/mol for B3LYP-D3/aug-cc-pVTZ and PBE0-D3/aug-cc-pVTZ, respectively.[97] HF-3c performs somewhat worse on the BIOMOLECULES set, viz., factor of 2.8 improvement over uncorrected HF-D3/MINIs.

Figure 2 shows that for all sets that comprise non-covalent binding energies, HF-D3/MINIs shows a strong bias towards over-binding, which is successfully corrected by ACPs and by the 3c correction. The standard deviation of the errors is also greatly reduced, except for S12L. The same cannot be said about the subsets composed of conformational energies (CONFORMERS group), for which uncorrected HF-D3/MINIs shows errors on both sides of the zero-average error line. The performance of HF-D3/MINIs-ACP for conformational energies is excellent, as demonstrated by the substantial decrease in MAE to 0.69 kcal/mol from 2.27 kcal/mol for uncorrected HF-D3/MINIs. By comparison, the MAE for the CONFORMERS group is 1.05 kcal/mol using HF-3c (Table 5). The results for the tripeptides (PCONF) and tetrapeptides (TRCONF) are particularly good, which is not surprising since our training set is dominated by peptide-peptide interactions (for instance, P26 contains PCONF). The decrease in MAE is also substantial for SCONF (sugar conformations, 5.20 to 1.17 kcal/mol) and BCONF (butane-1,4-diol, 2.40 to 0.50 kcal/mol), but smaller for MCONF (melatonin, 0.88 to 0.71 kcal/mol) and ACONF (hydrocarbons, 1.44 to 0.98 kcal/mol). In all CONFORMER subsets, the bias and the spread of the errors is reduced by the application of the ACPs and the 3c corrections.

Figure 2. Signed errors associated with the subsets of the validation set in Table 2 using HF-D3/MINIs with and without ACPs (HF-D3/MINIs-ACP) and the 3c correction approach (HF-3c). The circle represents the mean error (ME) and the error bar is the standard deviation of the error. The numbers on the right are mean absolute error (MAE) in kcal/mol.


MAE (kcal/mol)
HC12
HF-D3/MINIs 1.80
HF-D3/MINIs-ACP 0.26
HF-3c 0.42
ADIM6
HF-D3/MINIs 1.90
HF-D3/MINIs-ACP 0.21
HF-3c 0.47
CH4 · PAH
HF-D3/MINIs 1.19
HF-D3/MINIs-ACP 0.29
HF-3c 0.19
C2H4 · NT
HF-D3/MINIs 1.71
HF-D3/MINIs-ACP 0.67
HF-3c 0.48
CO2 · PAH
HF-D3/MINIs 1.64
HF-D3/MINIs-ACP 0.87
HF-3c 0.55
CO2 · NPHAC
HF-D3/MINIs 1.92
HF-D3/MINIs-ACP 0.89
HF-3c 0.63
S12L
HF-D3/MINIs 15.39
HF-D3/MINIs-ACP 10.27
HF-3c 6.28
S30L
HF-D3/MINIs 15.54
HF-D3/MINIs-ACP 7.53
HF-3c 6.01
SHIELDS38
HF-D3/MINIs 30.62
HF-D3/MINIs-ACP 4.99
HF-3c 7.67
IONICHB
HF-D3/MINIs 4.45
HF-D3/MINIs-ACP 2.47
HF-3c 2.68
SSI (charged)
HF-D3/MINIs 3.07
HF-D3/MINIs-ACP 1.87
HF-3c 2.38
A24
HF-D3/MINIs 0.73
HF-D3/MINIs-ACP 0.32
HF-3c 0.44
HSG
HF-D3/MINIs 1.69
HF-D3/MINIs-ACP 0.62
HF-3c 0.74
HBC6
HF-D3/MINIs 3.27
HF-D3/MINIs-ACP 0.76
HF-3c 1.13
ACONF
HF-D3/MINIs 1.44
HF-D3/MINIs-ACP 0.98
HF-3c 0.89
BCONF
HF-D3/MINIs 2.40
HF-D3/MINIs-ACP 0.50
HF-3c 0.58
MCONF
HF-D3/MINIs 0.88
HF-D3/MINIs-ACP 0.71
HF-3c 0.89
PCONF
HF-D3/MINIs 2.43
HF-D3/MINIs-ACP 0.50
HF-3c 2.28
SCONF
HF-D3/MINIs 5.20
HF-D3/MINIs-ACP 1.17
HF-3c 1.47
TRCONF
HF-D3/MINIs 5.14
HF-D3/MINIs-ACP 0.80
HF-3c 3.64
-20
-16
-12
-8
-6
-4
-2
0
2
4
6
8
Error distribution (kcal/mol)

4. Discussion and Outlook

The combined analysis of the HF-D3/MINIs-ACP performance on the training (Table 4 and Figure 1) and validation sets (Table 5 and Figure 2) offers some insight into the feasibility of using ACPs for developing a computationally inexpensive method based on minimal-basis-set quantum mechanical calculation. The overall performance is, in general, worse than conventional dispersion-corrected density functionals at the complete-basis-set limit and similar to the previously proposed HF-3c method. The performance of HF-D3/MINIs-ACP indicates that it is particularly suitable for biomolecules, and significantly better than uncorrected HF-D3/MINIs but with a similar computational cost. It is also clear that there is a certain degree of generality to the ACPs, since the MAEs for subsets of the validation set that have very little resemblance to the systems in the training set (e.g. the CO_2-CAPTURE or HYDROCARBONS groups) are consistently improved by the application of our ACPs. Nevertheless, the training set we utilized is dominated by systems derived from biological molecules, particularly proteins, and this is also reflected in the validation set. For instance, the errors in the TRCONF and PCONF conformational energies are much smaller than the other subsets. There is an essential limitation in the description of charged systems, however, caused by the very poor description of anions using a minimal-basis-set, which may have a negative impact on the calculation of zwitterionic species. For the same reason, strongly hydrogen-bonded systems (e.g. double hydrogen bonds in carboxylic acid dimers) are also difficult to model with a minimal-basis-set.

Another limitation of ACPs is that, in principle, they need to be developed for every atom in the system under study. However, work is under way to extend the training set to cover most atoms that usually appear in organic molecules, particularly P and S, which would enable the complete description of DNA and proteins using ACPs. Even if ACPs are only applied to a subset of the atoms in the system, their effect seems to reduce the error from HF-D3/MINIs-ACP, which is not surprising since it is dominated by the extreme basis-set incompleteness of the basis-set. For instance, we applied HF-D3/MINIs to the X40 set[99], comprising non-covalent binding energies of halogenated dimers. (Only the subset of molecules without Br and I was calculated, since there are no MINIs basis functions for those atoms.) Using HF-D3/MINIs, the MAE is 1.50 kcal/mol, which is reduced to 1.08 kcal/mol upon application of the ACPs, an MAE similar although slightly higher than HF-3c (0.94 kcal/mol). The error is reduced, even though all the molecules in X40 contain at least one halogen atom, for which no ACPs are available.

Our current training set is also somewhat skewed towards peptide-peptide interactions, and this is likely detrimental to the accurate modeling of other types of non-covalent forces. On the other hand, extending the training set is a relatively simple matter. A dispersion-corrected density functional (such as

LC-ωPBE-XDM) and a relatively large basis-set are good enough to generate reference data for our fits, since we have shown that these methods have in general a much higher accuracy than what we can obtain using an ACP-corrected minimal-basis-set HF calculation. The use of the LASSO fitting technique also allows training sets with hundreds of thousands or even millions of data points, which would not have been possible with the fitting procedure described in our previous work.[25]

Although extending the present work to create general-purpose ACPs is valuable, another positive feature of the current methodology is that it can be applied to develop ACPs for specific purposes. An example is our recently developed ACPs for water.[96] In addition, any property that is a linear mapping of the electronic energy can be targeted by the ACP, not just the total energy. This was nicely exemplified in our water ACPs[96], which indirectly brought the molecular dipole in water using BH&HLYP-D3/aug-cc-pVTZ to agreement with the experimental value to five significant digits. The ACPs presented in this work are inherently valuable as a general-purpose, computationally inexpensive exploratory tool, particularly for the purpose of modeling peptide-peptide interactions, which is very interesting in the field of quantum mechanical refinement of protein structures.[100,101]

References

(1) Risthaus, T.; Grimme, S. Benchmarking of London Dispersion-Accounting Density Functional Theory Methods on Very Large Molecular Complexes. *J. Chem. Theory Comput.* **2013**, *9* (3), 1580-1591.

(2) Ambrosetti, A.; Alfè, D.; DiStasio, Jr., R. A.; Tkatchenko, A. Hard Numbers for Large Molecules: Toward Exact Energetics for Supramolecular Systems. *J. Phys. Chem. Lett.* **2014**, *5* (5), 849−855.

(3) Otero-de-la-Roza, A.; Johnson, E. R. Predicting Energetics of Supramolecular Systems Using the XDM Dispersion Model. *J. Chem. Theory Comput.* **2015**, *11* (9), 4033–4040.

(4) Brandenburg, J. G.; Grimme, S. Dispersion Corrected Hartree–Fock and Density Functional Theory for Organic Crystal Structure Prediction. *Top Curr. Chem.* **2014**, *345*, 1-24.

(5) Whittleton, S. R.; Otero-de-la-Roza, A.; Johnson, E. R. The Exchange-Hole Dipole Dispersion Model for Accurate Energy Ranking in Molecular Crystal Structure Prediction. *J. Chem. Theory Comput.* **2017**, *13* (2), 441-450.

(6) Whittleton, S. R.; Otero-de-la-Roza, A.; Johnson, E. R. Exchange-Hole Dipole Dispersion Model for Accurate Energy Ranking in Molecular Crystal Structure Prediction II: Nonplanar Molecules. *J. Chem. Theory Comput.* **2017**, Article ASAP. (DOI: 10.1021/acs.jctc.7b00715)

(7) Thiel, W. Semiempirical Quantum-Chemical Methods. *WIREs Comput. Mol. Sci.* **2013**, *4* (2), 145–157.

(8) Brandenburg, J. G.; Hochheim, M.; Bredow, T.; Grimme, S. Low-Cost Quantum Chemical Methods for Noncovalent Interactions. *J. Phys. Chem. Lett.* **2014**, *5* (24), 4275–4284.

(9) Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chem. Rev.* **2016**, *116* (9), 5301–5337.

(10) Sure, R.; Brandenburg, J. G.; Grimme, S. Small Atomic Orbital Basis Set First-Principles Quantum Chemical Methods for Large Molecular and Periodic Systems: A Critical Analysis of Error Sources. *ChemistryOpen* **2016**, *5* (2), 94–109.

(11) Hohenstein, E. G.; David Sherrill, C. Wavefunction Methods for Noncovalent Interactions. *WIREs Comput. Mol. Sci.* **2011**, *2* (2), 304–326.

(12) Řezáč, J.; Hobza, P. Describing Noncovalent Interactions Beyond the Common Approximations: How Accurate Is the "Gold Standard," CCSD(T) At the Complete Basis Set Limit? *J. Chem. Theory Comput.* **2013**, *9* (5), 2151–2155.

(13) Riplinger, C.; Neese, F. An Efficient and Near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method. *J. Chem. Phys.* **2013**, *138* (3), 034106.

(14) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural Triple Excitations in Local Coupled Cluster Calculations with Pair Natural Orbitals. *J. Chem. Phys.* **2013**, *139* (13), 134101.

(15) Khaliullin, R. Z.; VandeVondele, J.; Hutter, J. Efficient Linear-Scaling Density Functional Theory for Molecular Systems. *J. Chem. Theory Comput.* **2013**, *9* (10), 4421-4427.

(16) Gale, J. D. SIESTA: A Linear-Scaling Method for Density Functional Calculations. In *Computational Methods for Large Systems: Electronic Structure Approaches for Biotechnology and Nanotechnology*; Reimers, J. R., Ed.; John Wiley & Sons, Inc: Hoboken, New Jersey, 2011; pp 45-75.

(17) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86). *J. Chem. Theory Comput.* **2017**, *13* (5), 1989–2009.

(18) Witte, J.; Neaton, J. B.; Head-Gordon, M.; Effective Empirical Corrections for Basis Set Superposition Error in the def2-SVPD Basis: gCP and DFT-C. *J. Chem. Phys.* **2017**, *146* (23), 234105.

(19) Krishnapriyan, A.; Yang, P.; Niklasson, A. M.N.; Cawkwell, M. J. Numerical Optimization of Density Functional Tight Binding Models: Application to Molecules Containing Carbon, Hydrogen, Nitrogen, and Oxygen. *J. Chem. Theory Comput.*, Just Accepted Manuscript. DOI: 10.1021/acs.jctc.7b00762

(20) DiLabio, G. A. Accurate Treatment of van Der Waals Interactions Using Standard Density Functional Theory Methods with Effective Core-Type Potentials: Application to Carbon-Containing Dimers. *Chem. Phys. Lett.* **2008**, *455* (4-6), 348–353.

(21) Mackie, I. D.; DiLabio, G. A. Interactions in Large, Polyaromatic Hydrocarbon Dimers: Application of Density Functional Theory with Dispersion Corrections. *J. Phys. Chem. A* **2008**, *112* (43), 10968–10976.

(22) Torres, E.; DiLabio, G. A. A (Nearly) Universally Applicable Method for Modeling Noncovalent Interactions Using B3LYP. *J. Phys. Chem. Lett.* **2012**, *3* (13), 1738–1744.

(23) DiLabio, G. A.; Koleini, M. Dispersion-Correcting Potentials Can Significantly Improve the Bond Dissociation Enthalpies and Noncovalent Binding Energies Predicted by Density-Functional Theory. *J. Chem. Phys.* **2014**, *140* (18), 18A542.

(24) van Santen, J. A.; DiLabio, G. A. Dispersion Corrections Improve the Accuracy of Both Noncovalent and Covalent Interactions Energies Predicted by a Density-Functional Theory Approximation. *J. Phys. Chem. A* **2015**, *119* (25), 6703–6713.

(25) Otero-de-la-Roza, A.; DiLabio, G. A. Transferable Atom-Centered Potentials for the Correction of Basis Set Incompleteness Errors in Density-Functional Theory. *J. Chem. Theory Comput.* **2017**, *13* (8), 3505-3524.

(26) Perdew, J. P.; Levy, M. Physical Content of the Exact Kohn-Sham Orbital Energies: Band Gaps and Derivative Discontinuities. *Phys. Rev. Lett.* **1983**, *51* (20), 1884–1887.

(27) Godby, R. W.; Schlüter, M.; Sham, L. J. Self-Energy Operators and Exchange-Correlation Potentials in Semiconductors. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37* (17), 10159–10175.

(28) Seidl, A.; Görling, A.; Vogl, P.; Majewski, J. A.; Levy, M. Generalized Kohn-Sham Schemes and the Band-Gap Problem. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *53* (7), 3764–3774.

(29) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Insights into Current Limitations of Density Functional Theory. *Science* **2008**, *321* (5890), 792–794.

(30) Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Localization and Delocalization Errors in Density Functional Theory and Implications for Band-Gap Prediction. *Phys. Rev. Lett.* **2008**, *100* (14), 146401.

(31) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic-Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098.

(32) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37* (2), 785.

(33) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. Results Obtained with the Correlation-Energy Density Functionals of Becke and Lee, Yang and Parr. *Chem. Phys. Lett.* **1989**, *157* (3), 200-206.

(34) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865.

(35) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648.

(36) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98* (45), 11623-11627.

(37) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular

Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107* (13), 3902-3909.

(38) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods I. Method. *J. Comput. Chem.* **1989**, *10* (2), 209-220.

(39) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model* **2007**, *13* (12), 1173-1123.

(40) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J. Mol. Model* **2013**, *19* (1), 1-32.

(41) McNamara, J. P.; Hillier, I. H. Semi-Empirical Molecular Orbital Methods Including Dispersion Corrections for the Accurate Prediction of the Full Range of Intermolecular Interactions in Biomolecules. *Phys. Chem. Chem. Phys.* **2007**, *9* (19), 2362-2370.

(42) Tuttle, T.; Thiel, W. OMx-D: Semiempirical Methods with Orthogonalization and Dispersion Corrections. Implementation and Biochemical Application. *Phys. Chem. Chem. Phys.* **2008**, *10* (16), 2159–2166.

(43) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza P. Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes. *J. Chem. Theory Comput.* **2009**, *5* (7), 1749–1760.

(44) Řezáč, J.; Hobza, P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *J. Chem. Theory Comput.* **2012**, *8* (1), 141–151.

(45) Sure, R.; Grimme, S. Corrected Small Basis Set Hartree-Fock Method for Large Systems. *J. Comput. Chem.* **2013**, *34* (19), 1672–1685.

(46) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104.

(47) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1456–1465.

(48) Johnson, E. R.; Becke, A. D. A Post-Hartree-Fock Model of Intermolecular Interactions: Inclusion of Higher-Order Corrections. *J. Chem. Phys.* **2006**, *124* (17), 174104.

(49) Kruse, H.; Grimme, S. A Geometrical Correction for the Inter- and Intra-Molecular Basis Set Superposition Error in Hartree-Fock and Density Functional Theory Calculations for Large Systems. *J. Chem. Phys.* **2012**, *136* (15), 154101.

(50) Reilly, A. M. *et* al. Report on the Sixth Blind Test of Organic Crystal Structure Prediction Methods. Acta Cryst. **2016**, *B72*, 439−459.

(51) Kahn, L. R.; Baybutt, P.; Truhlar, D. G. Ab Initio Effective Core Potentials: Reduction of All-electron Molecular Structure Calculations to Calculations Involving Only Valence Electrons. *J. Chem. Phys.* **1976**, *65* (10), 3826–3853.

(52) Christiansen, P. A.; Lee, Y. S.; Pitzer, K. S. Improved Ab Initio Effective Core Potentials for Molecular Calculations. In *World Scientific Series in 20th Century Chemistry, Molecular Structure and Statistical Thermodynamics: Selected Papers of Kenneth S Pitzer;* Pitzer, K. S., Ed.; WORLD SCIENTIFC PUB CO INC, **1993**; Vol. 1; pp 147–152.

(53) Christiansen, P. A.; Lee, Y. S.; Pitzer, K. S. Improved Ab Initio Effective Core Potentials for Molecular Calculations. *J. Chem. Phys.* **1979**, *71* (11), 4445–4450.

(54) Tatewaki, H.; Huzinaga, S. A Systematic Preparation of New Contracted Gaussian-Type Orbital Sets. III. Second-Row Atoms from Li through Ne. *J. Comput. Chem.* **1980**, *1* (3), 205–228.

(55) Otero-de-la-Roza, A.; Johnson, E. R. Non-Covalent Interactions and Thermochemistry Using XDM-Corrected Hybrid and Range-Separated Hybrid Density Functionals. *J. Chem. Phys.* **2013**, *138* (20), 204109.

(56) Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Statist. Soc. B* **1996**, *58* (1), 267-288.

(57) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; calmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, Revision D.01; Gaussian Inc.: Wallingford, CT, 2009.

(58) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. Benchmark Database of Accurate (MP2 and CCSD(T) Complete Basis Set Limit) Interaction Energies of Small Model Complexes, DNA Base Pairs, and Amino Acid Pairs. *Phys. Chem. Chem. Phys.* **2006**, *8* (17), 1985–1993.

(59) Gráfová, L.; Pitoňák, M.; Řezáč, J.; Hobza, P. Comparative Study of Selected Wave Function and Density Functional Methods for Noncovalent Interaction Energy Calculations Using the Extended S22 Data Set. *J. Chem. Theory Comput.* **2010**, *6* (8), 2365–2376.

(60) Rezáč, J.; Riley, K. E.; Hobza, P. S66: A Well-Balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J. Chem. Theory Comput.* **2011**, *7* (8), 2427–2438.

(61) Řezáč, J.; Riley, K. E.; Hobza, P. Extensions of the S66 Data Set: More Accurate Interaction Energies and Angular-Displaced Nonequilibrium Geometries. *J. Chem. Theory Comput.* **2011**, *7* (11), 3466–3470.

(62) Brauer, B.; Kesharwani, M. K.; Kozuch, S.; Martin, J. M. L. The S66x8 Benchmark for Noncovalent Interactions Revisited: Explicitly Correlated Ab Initio Methods and Density Functional Theory. *Phys. Chem. Chem. Phys.* **2016**, *18* (31), 20905–20925.

(63) Smith, D. G. A.; Burns, L. A.; Patkowski, K.; Sherrill, C. D. Revised Damping Parameters for the D3 Dispersion Correction to Density Functional Theory. *J. Phys. Chem. Lett.* **2016**, *7* (12), 2197–2203.

(64) Parker, T. M.; Sherrill, C. D. Assessment of Empirical Models versus High-Accuracy Ab Initio Methods for Nucleobase Stacking: Evaluating the Importance of Charge Penetration. *J. Chem. Theory Comput.* **2015**, *11* (9), 4197–4204.

(65) Burns, L. A.; Faver, J. C.; Zheng, Z.; Marshall, M. S.; Smith, D. G. A.; Vanommeslaeghe, K.; MacKerell, A. D.; Merz, K. M.; Sherrill, C. D. The BioFragment Database (BFDb): An Open-Data Platform for Computational Chemistry Analysis of Noncovalent Interactions. *J. Chem. Phys.* **2017**, *147*(16), 161727.

(66) Valdes, H.; Pluhácková, K.; Pitonák, M.; Rezác, J.; Hobza, P. Benchmark Database on Isolated Small Peptides Containing an Aromatic Side Chain: Comparison between Wave Function and Density Functional Theory Methods and Empirical Force Field. *Phys. Chem. Chem. Phys.* **2008**, *10* (19), 2747–2757.

(67) Goerigk, L.; Grimme, S. A Thorough Benchmark of Density Functional Methods for General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions. *Phys. Chem. Chem. Phys.* **2011**, *13* (14), 6670–6688.

(68) Gruzman, D.; Karton, A.; Martin, J. M. L. Performance of Ab Initio and Density Functional Methods for Conformational Equilibria ofC_nH_{2n+2} Alkane Isomers (n=4−8). *J. Phys. Chem. A* **2009**, *113* (43), 11974–11983.

(69) Reha, D.; Valdés, H.; Vondrásek, J.; Hobza, P.; Abu-Riziq, A.; Crews, B.; de Vries, M. S. Structure and IR Spectrum of Phenylalanyl-Glycyl-Glycine Tripeptide in the Gas-Phase: IR/UV Experiments, Ab Initio Quantum Chemical Calculations, and Molecular Dynamic Simulations. *Chem. Eur. J.* **2005**, *11* (23), 6803–6817.

(70) Csonka, G. I.; French, A. D.; Johnson, G. P.; Stortz, C. A. Evaluation of Density Functionals and Basis Sets for Carbohydrates. *J. Chem. Theory Comput.* **2009**, *5* (4), 679–692.

(71) Case, D. A.; Betz, R. M.; Cerutti, D. S.; Cheatham, T. E.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T.S.; LeGrand, S.; Li, P.; Lin, C.; Luchko, T.; Luo, R.; Madej, B.; Mermelstein, D.; Merz, K. M.; Monard, G.; Nguyen, H.; Nguyen, H. T.; Omelyan, I.; Onufriev, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Simmerling, C. L.; Botello-Smith, W. M.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Xiao, L.; Kollman, P. A. AMBER 2016, University of California, San Francisco, 2016.

(72) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713.

(73) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins: Struct. Funct. Bioinf.* **2006**, *65* (3), 712–725.

(74) Osborne, M. R.; Presnell, B.; Turlach, B. A. A New Approach to Variable Selection in Least Squares Problems. *IMA Journal of Numerical Analysis* **2000**, *20* (3), 389-403.

(75) Schmidt, M. Graphical Model Structure Learning with L1-Regularization. Ph.D. Thesis, The University of British Columbia, Vancouver, August 2010.

(76) Schmidt, M.; Fung, G.; Rosales, R. *Optimization Methods for L1-Regularization*; Technical Report for The University of British Columbia (TR-2009-19), August 2009.

(77) Granatier, J.; Pitoňák, M.; Hobza, P. Accuracy of Several Wave Function and Density Functional Theory Methods for Description of Noncovalent Interaction of Saturated and Unsaturated Hydrocarbon Dimers. *J. Chem. Theory Comput.* **2012**, *8* (7), 2282–2292.

(78) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M. Estimated MP2 and CCSD(T) Interaction Energies of N-Alkane Dimers at the Basis Set Limit: Comparison of the Methods of Helgaker *et al.* and Feller. *J. Chem. Phys.* **2006**, *124* (11), 114304.

(79) Smith, D. G. A.; Patkowski, K. Toward an Accurate Description of Methane Physisorption on Carbon Nanotubes. *J. Phys. Chem. C* **2014**, *118* (1), 544–550.

(80) Smith, D. G. A.; Patkowski, K. Interactions between Methane and Polycyclic Aromatic Hydrocarbons: A High Accuracy Benchmark Study. *J. Chem. Theory Comput.* **2013**, *9* (1), 370–389.

(81) Smith, D. G. A.; Patkowski, K. Benchmarking the CO_2 Adsorption Energy on Carbon Nanotubes. *J. Phys. Chem. C* **2015**, *119* (9), 4934–4948.

(82) Li, S.; Smith, D. G. A.; Patkowski, K. An Accurate Benchmark Description of the Interactions between Carbon Dioxide and Polyheterocyclic Aromatic Compounds Containing Nitrogen. *Phys. Chem. Chem. Phys.* **2015**, *17* (25), 16560–16574.

(83) Grimme, S. Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory. *Chem. Eur. J.* **2012**, *18* (32), 9955–9964.

(84) Antony, J.; Sure, R.; Grimme, S. Using Dispersion-Corrected Density Functional Theory to Understand Supramolecular Binding Thermodynamics. *Chem. Commun.* **2015**, *51* (10), 1764-1774.

(85) Sure, R.; Grimme, S. Correction to Comprehensive Benchmark of Association (Free) Energies of Realistic Host–Guest Complexes. *J. Chem. Theory Comput.* **2015**, *11* (12), 5990–5990.

(86) Sure, R.; Grimme, S. Comprehensive Benchmark of Association (Free) Energies of Realistic Host-Guest Complexes. *J. Chem. Theory Comput.* **2015**, *11* (8), 3785–3801.

(87) Temelso, B.; Archer, K. A.; Shields, G. C. Benchmark Structures and Binding Energies of Small Water Clusters with Anharmonicity Corrections. *J. Phys. Chem. A* **2011**, *115* (43), 12034–12046.

(88) Marshall, M. S.; Burns, L. A.; Sherrill, C. D. Basis Set Convergence of the Coupled-Cluster Correction, δ(MP2)(CCSD(T)): Best Practices for Benchmarking Non-Covalent Interactions and the Attendant Revision of the S22, NBC10, HBC6, and HSG Databases. *J. Chem. Phys.* **2011**, *135* (19), 194102.

(89) Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; David Sherrill, C.; Merz, K. M. Formal Estimation of Errors in Computed Absolute Interaction Energies of Protein−Ligand Complexes. *J. Chem. Theory Comput.* **2011**, *7* (3), 790–797.

(90) Thanthiriwatte, K. S.; Hohenstein, E. G.; Burns, L. A.; Sherrill, C. D. Assessment of the Performance of DFT and DFT-D Methods for Describing Distance Dependence of Hydrogen-Bonded Interactions. *J. Chem. Theory Comput.* **2011**, *7* (1), 88–96.

(91) Kozuch, S.; Bachrach, S. M.; Martin, J. M. L. Conformational Equilibria in Butane-1,4-diol: A Benchmark of a Prototypical System with Strong Intramolecular H-bonds. *J. Phys. Chem. A* **2014**, *118* (1), 293−303.

(92) Fogueri, U. R.; Kozuch, S.; Karton, A.; Martin, J. M. L. The Melatonin Conformer Space: Benchmark and Assessment of Wave Function and DFT Methods for a Paradigmatic Biological and Pharmacological Molecule. *J. Phys. Chem. A* **2013**, *117* (10), 2269–2277.

(93) Goerigk, L.; Karton, A.; Martin, J. M. L.; Radom, L. Accurate Quantum Chemical Energies for Tetrapeptide Conformations: Why MP2 Data with an Insufficient Basis Set Should be Handled with Caution. *Phys. Chem. Chem. Phys.* **2013**, *15* (19), 7028-7031.

(94) Li. A.; Muddana, H. S.; Gilson, M. K. Quantum Mechanical Calculation of Noncovalent Interactions: A Large-Scale Evaluation of PMx, DFT, and SAPT Approaches. *J. Chem. Theory Comput.* **2014**, *10* (4), 1563−1575.

(95) Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Cryst.* **1976**, A32, 922-923.

(96) Holmes, J. D.; Otero-de-la-Roza, A.; DiLabio, G. A. Accurate Modeling of Water Clusters with Density-Functional Theory Using Atom-Centered Potentials. *J. Chem. Theory Comput.* **2017**, *13* (9), 4205-4215.

(97) Burns, L. A.; Vázquez-Mayagoitia, A.; Sumpter, B. G.; Sherrill, C. D. Density-Functional Approaches to Noncovalent Interactions: A Comparison of Dispersion Corrections (DFT-D), Exchange-Hole Dipole Moment (XDM) Theory, and Specialized Functionals. *J. Chem. Phys.* **2011**, *134* (8), 084107.

(98) DiLabio, G. A.; Otero-de-la-Roza, A. Noncovalent Interactions in Density Functional Theory. In *Reviews in Computational Chemistry*; Parill, A. L., Lipkowitz, K. B., Eds.; John Wiley & Sons, Inc: Hoboken, New Jersey, 2016; Vol. 29; pp 1–97.

(99) Řezáč, J.; Riley, K. E.; Hobza, P. Benchmark Calculations of Noncovalent Interactions of Halogenated Molecules. *J. Chem. Theory Comput.*, **2012**, *8* (11), 4285-4292.

(100) Goerigk, L.; Reimers, J. R. Efficient Methods for the Quantum Chemical Treatment of Protein Structures: The Effects of London-Dispersion and Basis-Set Incompleteness on Peptide and Water-Cluster Geometries. *J. Chem. Theory Comput.* **2013**, *9* (7), 3240–3251.

(101) Goerigk, L.; Collyer, C. A.; Reimers, J. R. Recommending Hartree−Fock Theory with London-Dispersion and Basis-Set-Superposition Corrections for the Optimization or Quantum Refinement of Protein Structures. *J. Phys. Chem. B* **2014**, *118* (50), 14612−14626.

Part III

The results from Chapter 3 suggested that atom-centered potentials (ACPs) could be designed to correct the errors of quantum mechanical methods for target molecular properties by fitting them to appropriate reference data. To develop new ACPs that can be applied to a wide range of systems and more molecular properties, there was a need to assemble a training set that contained more diverse data points than used in Chapter 3.

The scarcity of certain reference data led to the development of four new data sets, as detailed in Chapters 4–6. In Chapter 4, we undertook the generation of the PEPCONF data set, which contains information about structures and conformational energies (a non-covalent property) of a diverse collection of peptide-like systems. The work in Chapter 5 details the development of a large data set of bond separation energies. Chapter 6 describes the development of a data set of reaction energies and reaction barrier heights.

While these data sets were generated specifically for ACP development, they will find use in other work, e.g., assessment of density-functional theory methods. As such, the data sets represent valuable additions to the literature.

The supporting information associated with Chapters 4 and 6 is provided in Appendices 2 and 3 of this dissertation, respectively. Note that there is no supporting information for Chapter 5. Other supporting files have also been deposited to the figshare repository and are openly available at the following URL/DOI: https://doi.org/10.6084/m9.figshare.16912201. The references of the published paper are as follows: (i) (PEPCONF) Prasad, V. K.; Otero-de-la-Roza, A.; DiLabio, G. A. *Sci. Data* 2019, 6, 180310. © Copyright 2019 Springer Nature Limited. (DOI: 10.1038/sdata.2018.310), (ii) (BSE49) Prasad, V. K.; Khalilian, M. H.; Otero-de-la-Roza, A.; DiLabio, G. A. *Sci. Data* 2021, 8, 300. © Copyright 2021 Springer Nature Limited. (DOI: 10.1038/s41597-021-01088-2), and (iii) (BH9) Prasad, V. K.; Pei, Z.; Edelmann, S; Otero-de-la-Roza, A.; DiLabio, G. A. *J. Chem. Theory Comput.* 2021. © Copyright 2021 American Chemical Society. (DOI: 10.1021/acs.jctc.1c00694)

Chapter 4

PEPCONF, a diverse data set of peptide conformational energies

Abstract

We present an extensive and diverse database of peptide conformational energies. Our database contains five different classes of model geometries: dipeptides, tripeptides, and disulfide-bridged, bioactive, and cyclic peptides. In total, the database consists of 3775 conformational energy data points and 4530 conformer geometries. All the reference energies have been calculated at the LC-ωPBE-XDM/aug-cc-pVTZ level of theory, which is shown to yield conformational energies with an accuracy in the order of tenths of a kcal/mol when compared to complete basis set, coupled-cluster reference data. The peptide conformational data set (PEPCONF) is presented as a high-quality reference data for the development and benchmarking studies of molecular-mechanics and semi-empirical electronic structure methods, which are the most commonly used techniques in the modeling of medium to large proteins.

1. Background & Summary

The structure and function of proteins are governed by the intermolecular interactions between their building blocks, amino acids. The accurate prediction of protein folding and ligand binding energetics depends on how well the atomistic computational modeling method employed captures the interactions between individual amino acids. For this reason, the results obtained from the computational methods commonly employed to model proteins, such as force field and semi-empirical electronic structure methods, are usually compared to, and parametrized against, those obtained from higher-level computational methods. A database of peptide conformational energies is an ideal benchmark set for testing and parameterizing computational methods since conformational energies capture the interplay between bonded and non-bonded interactions that are present in proteins.

Similar sets to the one proposed in this work are available in the literature, but they tend to be small and focus on specific peptide interactions or otherwise focus exclusively on single amino acids. In 2008, Hobza and co-workers presented a benchmark database of conformational energies for a set of 76 conformers of four tripeptides and a dipeptide containing aromatic side chains.[1] The conformational energies were calculated at the CCSD(T)/complete-basis-set (CBS) level of theory and, in the same work, were used to assess lower-level quantum-mechanical (QM) methods. The reference data for a subset of Hobza's set (named PCONF) was updated by Smith and co-workers[2], and later by Goerigk and co-workers[3]. Wilke *et al.* proposed a set of conformational energies for cysteine known as CYCONF[4], eight conformational energies of tetrapeptide conformers were proposed by Goerigk *et al.*[5], and Ropo *et al.*

presented a conformer data set of capped and uncapped versions of proteinogenic amino acids and their interactions with divalent cations evaluated at ‘PBE+vdW’ level of theory[6]. More recently, Martin and co-workers re-optimized the conformer structures of twenty proteinogenic amino acids from a previously published set by Yuan, Mills, Popelier, and Jensen (the YMPJ database).[7,8] These structures were then used to generate a new conformational energy database of isolated amino acid monomers containing 466 data points. A database of macrocyclic conformers, called MPCONF196, has recently been published.[9] The MPCONF196 set contains conformational energies of eight macrocyclic compounds including cyclic peptides of varying sizes. To our knowledge, MPCONF196 is the only set in the literature that considers cyclic peptides. Several of the data sets described above have been compiled into supersets. Hobza’s 2008 data set was included as a subset of the MPCONF196 benchmark database.[1,9] Similarly, the CYCONF, PCONF, TPCONF, and YMPJ sets of conformational energies were incorporated in the GMTKN databases by Grimme and co-workers.[3,10,11]

To best of our knowledge, an extensive database of polypeptide conformations is not yet available in the literature. It is likely that the absence of a comprehensive data set rests on the fact that structural complexity and the computational cost of obtaining reference-quality data increases with system size. A comprehensive set of data that contains reference conformational energies on a diversity of small peptides would provide valuable information to those engaged in the development of atomistic computational methods for protein modeling. Producing such a database of conformational energies of diverse polypeptides would ensure a uniform high-quality standard in the reference data by eliminating the need to collect and verify data gathered from various sources, which may differ substantially in their mode of generation and quality.

In this work, we have undertaken a substantial computational effort to generate a large, comprehensive polypeptide conformational energy data set using dispersion-corrected range-separated density-functional theory. The data set has several important features: 1) The conformational energies were obtained using a single computational method, which results in data with uniform quality; 2) The quality of the results obtained from the computational method we used to obtain the conformational energies is benchmarked against those obtained using complete basis set, coupled-cluster methods. This provides a means for assessing the quality of our database; 3) The computational method we used to obtain conformational energies is of much higher quality than conventional force field methods used for large-scale protein modeling and is therefore fit for testing and parametrization of conventional force field methods. Therefore, our data can be used for molecular mechanics force field development[12–14], and parametrization of cost-effective computational procedures like Atom-Centered Potentials (ACP)[15,16] and other low-cost correction approaches[17–19]. It also serves as a direct source for comparative benchmark studies of various energy functions[20–27], semi-empirical approaches[28–40], and inexpensive electronic structure methods[41–47] in the context of protein modeling.

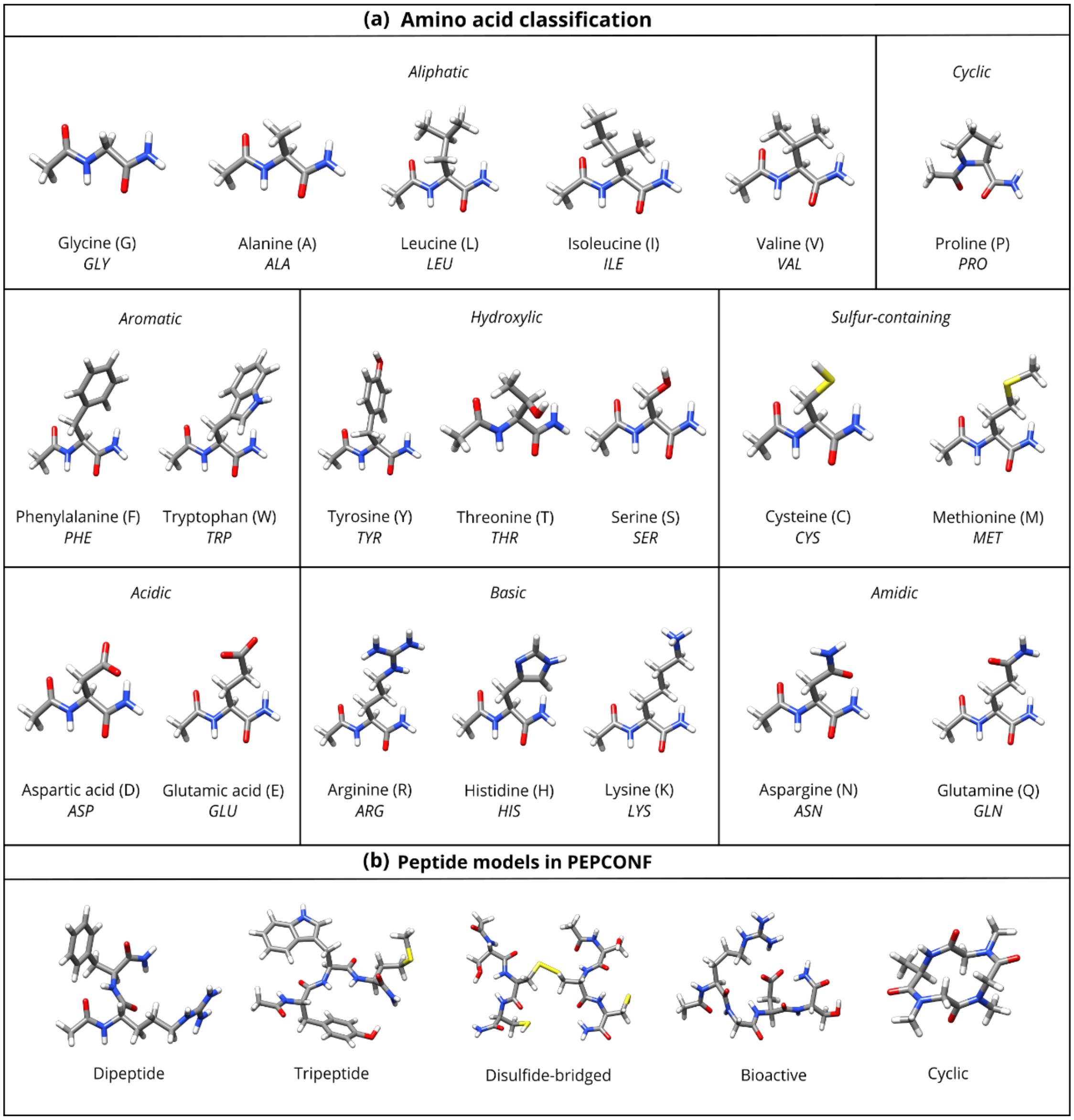


Figure 1. Molecular structure of the amino acids and representative peptide model systems considered in this work. (a) The classification of the twenty standard proteinogenic amino acids by the nature of their side-chains. The N-terminal and C-terminal are capped with acetyl and primary amide group, respectively. The single- and three-letter codes for each amino acid are also provided. (b) A representative candidate from each of the five different classes of peptide model systems considered in the PEPCONF data set.

2. Methods

Generation of the model geometries

The PEPCONF set comprises five different kinds of model systems:

- Dipeptides: all unique pairs of the twenty standard proteinogenic amino acids were selected (for instance, ALA-GLY and GLY-ALA were considered to be the same from the perspective of side chain-side chain interactions), leading to 136 neutral and 74 charged dipeptide geometries.
- Tripeptides: unique combinations of tripeptide sequences were selected similarly but, in order to limit the number of combinations, one representative amino acid was chosen from each of the side-chain categories in Figure 1a: Leucine for aliphatic, Proline for cyclic, Tryptophan for aromatic, Tyrosine for hydroxylic, Methionine for sulfur-containing, neutral Glutamic acid for acidic, Histidine for basic, and Glutamine for amidic side-chains. This yielded a total of 288 unique combinations of amino acid trimers.
- Disulfide-bridged: oligopeptides where the two cysteine residues are internally connected via a disulfide bond (154 model systems).
- Bioactive: oligopeptides where the chosen residue sequences were found to be associated with bio-functionality as reported in the literature[48] (39 model systems).
- Cyclic: oligopeptides where the N-terminus and C-terminus of the peptide backbone are connected to form a circular bond (64 model systems).

Structures

The initial gas-phase model geometries of the dipeptides, tripeptides, and bioactive peptides were generated using the *sequence* command in the *tleap* tool of *Amber16* software package[49–51]. The disulfide-bridged and cyclic peptides were generated manually from structures taken from the *Protein Data Bank* (PDB)[52,53] and the *Cambridge Structural Database* (CSD)[54,55], respectively. The N-terminal(s) and C-terminal(s) of all the representative model structures except for cyclic peptides were capped with acetyl (ACE) and primary amide (NHE) groups, respectively. The complete list of all the peptide structures considered in this work is provided in the supplementary file accompanying this article (Supplementary File 1).

The initial model geometries of disulfide-bridged oligopeptides were generated using an in-house fragmentation code and a combination of various *Amber16* tools like *pdb4amber*, *tleap*, and *pytleap*. Representative structures were initially obtained from searches of the *Protein Data Bank* (PDB) using the online advanced search interface with the following criteria: (i) only one disulfide bond, (ii) X-ray resolution between 2.5-3.5 Å, (iii) no modified polymeric residues, (iv) no free ligands, and (v) representative structures at 100% sequence identity. The resulting 191 hits were then processed with the *pdb4amber* tool to remove the water molecules from the PDB files and to select the most populous conformer. We then discarded 37 out of the 191 clean PDB files because the most populated conformer did not contain a disulfide bond. Finally, the clean PDB files were truncated using our fragmentation code and the disulfide-bridged cysteine residues of each model system were extracted along with at most four neighboring backbone residues. Each system was manually checked and then processed with *pytleap* and *tleap* to add the missing hydrogen atoms and terminal capping groups.

The initial model geometries of cyclic peptides were found using the *Conquest* software package to search for crystal structures in the *Cambridge Structural Database*. Cyclic sequences of proteinogenic amino acids were searched using the peptide building query tool. The following search criteria were used: (i) 3D coordinates must have been determined, (ii) R-factor less than or equal to 0.05, (iii) only non-disordered crystals, (iv) no errors present, (v) no ions present. The resulting structures were then exported to 'mol2' files which were converted to 'xyz' format using *Openbabel*[56,57] and loaded in the *Avogadro*[58,59] software package for visual inspection. Structures without a proper cyclic peptide backbone were not considered. Finally, the missing H-atoms were added using *Avogadro*.

The initial geometries of all the model systems, with the exception of cyclic peptides, were subjected to *Amber ff14SB*[21] unconstrained force field energy relaxations using the *sander* module of *Amber16*.

Conformational search

A force field-based high-temperature molecular dynamics (HTMD) simulation approach[60] was used in a manner similar to previous studies in the literature[61–64] to generate the conformers for the non-cyclic peptides. Initial structures were subjected to canonical ensemble simulations with Langevin dynamics scaling at a temperature of 900 K. The MD steps were performed with the *sander* module of *Amber16* without solvent or periodicity. A heating (equilibration) step of 200 picoseconds was followed by a production run of 4.2 nanoseconds. Structures along the trajectory of the production run were sampled at uniform time intervals, resulting in 4000 conformers for each peptide model system. Each conformer was subjected to energy minimization using the *Amber ff14SB* force field.

The *Amber ff14SB* force field does not contain parameter for cyclic peptides. We therefore used the *RDKit* software package[65] to generate cyclic peptide conformations. The accuracy and speed of *RDKit's* conformer generation approach in comparison to other freely available conformer generation toolkits was reviewed in Ref. 66, where it was reported that the program is suited for less flexible molecules like the cyclic peptides considered in this work. A distance-geometry-based stochastic method[67] was used to yield 100 conformers for each cyclic peptide. A very similar approach was recently used to generate the 3D conformations reported in the ANI-1 data set.[68]

Conformer binning strategy

The list of relaxed conformers was pruned using a binning strategy. Each set of non-cyclic conformers was sorted according to the force field energy, from most to least stable. The least stable conformers that populate the upper half of the list were removed, and the remainder of the list was divided into thirty equal energy intervals. From each interval, one conformer geometry was selected and was subjected to a single-point energy calculation with the BLYP gradient-corrected density functional[69,70], and the 6-31G* basis set[71,72], combined with Grimme's D3 dispersion-correction method[73,74] with Becke-Johnson (BJ) damping function[75–81] and recently developed basis set incompleteness potentials (BSIP)[82]. The calculations with the BLYP-D3(BJ)/6-31G*-BSIP level of theory were carried out using the *Gaussian* software package[83,84], with SCF convergence criterion of 10^{-6} Hartrees and pruned integration grid with 99 radial and 590 angular points (ultrafine grid). The resulting BLYP-D3(BJ)/6-31G*-BSIP energies were used to select the six most stable conformers out of the thirty for entry into the PEPCONF data set.

In the case of the cyclic peptides, the 100 conformers generated by *RDKit* were geometry-optimized at the BLYP-D3(BJ)/MINIs-BSIP[69,70,73-82,85] level of theory using the *Gaussian* package. The calculations employed SCF convergence criterion of 10^{-8} Hartrees, ultrafine integration grid, and the default optimization convergence criteria (maximum force=$4.5\text{x}10^{-4}$ Hartrees/Bohr, RMS force=$3\text{x}10^{-4}$ Hartrees/Bohr, maximum displacement=$1.8\text{x}10^{-3}$ Bohr, RMS displacement=$1.2\text{x}10^{-3}$ Bohr). The equilibrium geometries were sorted by energy and six conformations from equally-spaced energy intervals covering the whole energy range were then selected. The six conformations were then subjected to further geometry optimizations using BLYP-D3(BJ)/6-31G*-BSIP with the same SCF and grid settings as above and a '*verytight'* optimization convergence criteria (maximum force=$2\text{x}10^{-6}$ Hartrees/Bohr, RMS force=$1\text{x}10^{-6}$ Hartrees/Bohr, maximum displacement=$6\text{x}10^{-6}$ Bohr, RMS displacement=$4\text{x}10^{-6}$ Bohr).

Generation of the reference energies

The PEPCONF data set contains 5 relative conformational energies (from the 6 conformations) for each peptide model system considered, yielding a total of 3775 data points from 4530 conformer structures. The reference energies were calculated with the LC-ωPBE[86,87] range-separated density functional, and the aug-cc-pVTZ basis set of Dunning and co-workers[88–90], combined with the exchange-hole dipole moment

(XDM) dispersion-correction technique[75–81]. The rationale for this choice is that it offers a good compromise between accuracy and speed, and we expect range-separated hybrid functionals to minimize the impact of functional delocalization error on zwitterionic and charged species.[91] The resulting DFT-based approach was chosen as the reference level because of its excellent performance for gas-phase results of relative conformational energies (see Technical Validation).

A wave-function based approach like the "gold-standard" CCSD(T)/CBS would provide more reliable relative conformer energies.[92,93] However, CCSD(T)/CBS calculations are not feasible for systems (23-166 number of atoms) included in the data set. In addition, the PEPCONF data set is intended as a database for parametrization and benchmarking of force fields, semi-empirical methods and other low computational cost methods, which have much higher errors in conformational energies than those associated with LC-ωPBE-XDM/aug-cc-pVTZ. Future revisions of the PEPCONF data will be possible as computing power increases and approximate but accurate CCSD(T) methods are developed[94,95].

3. Code Availability

The molecular dynamics simulations were carried out using *Amber16*, which is available from http://ambermd.org/ through a commercial license. The *Amber16* tools *pdb4amber*, *tleap*, and *pytleap* used for peptide structure editing and manipulation are part of the *Amber16* software package. The *Cambridge Structural Database 2018* and the *Conquest* program are distributed under a commercial license at https://www.ccdc.cam.ac.uk/. *RDKit* is an open-source cheminformatics software made available under the Berkeley Software Distribution (BSD) license at https://www.rdkit.org/. The *OpenBabel* software package was used for file-type interconversions and is freely available from http://openbabel.org/ under the GPL license. The *Avogadro* molecular editor and visualizer is an open-source program available at https://avogadro.cc/. The quantum-mechanical calculations were performed using the *Gaussian09/16* software packages, which can be purchased from Gaussian Inc. (http://gaussian.com/) under a commercial license. Finally, the Basis-Set Incompleteness Potentials (BSIP) for BLYP-D3(BJ)/MINIs and BLYP-D3(BJ)/6-31G* level of theory can be obtained from the Supporting Information of Ref. 82.

4. Data Records

The conformational reference energies (in kcal/mol) and coordinates (in Å) of the conformer geometries present in the PEPCONF data set are publicly available free-of-charge from the Figshare (Data Citation 1) and GitHub (https://github.com/aoterodelaroza/pepconf) repositories in the plain-text DB-format described in Table 1. The atomic coordinates of the conformer geometries are also stored in a plain-text XYZ-format. The PEPCONF set contains five DB-format and six XYZ-format files for each peptide model system. In total, deposited files include 3775 DB-format files and 4530 XYZ-format files stored in their respective peptide classification directory named Dipeptide, Tripeptide, Disulfide, Bioactive, and Cyclic. A CSV-format file is also provided in each directory and contains the reference energy values for all the peptide systems in that directory.

5. File Format

For each molecule, the reference conformational energy, relative to the lowest-energy structure, and the atomic coordinates are stored in a file named *MoleculeName_A.db*, where A is the conformer identification number (1-5, ordered from lowest to highest relative energy). The Cartesian coordinates of the atoms are stored in files named *MoleculeName_B.xyz,* where B is 0-5 (ordered from lowest to highest relative energy), with 0 representing the lowest-energy reference structure.

The DB-format file contains a header line specifying the reference energy value (in kcal/mol) followed by two '*molc'* (short for molecule) blocks containing a unique integer identifier, charge, multiplicity, and the atomic coordinates (in Å) of the peptide conformer and its corresponding lowest energy conformer. The XYZ-format file contains a header line defining the number of atoms N, a comment line containing the charge and multiplicity, and N lines with each containing element type and X, Y, Z coordinates (in Å). The CSV-format file is a comma-separated plain-text file containing multiple lines and three columns. The columns are: (i) identification number, (ii) name of the peptide, and (iii) reference conformational energy (in kcal/mol).

Table 1. A description of the DB-format file or the database-file format (.db) for a peptide system containing N number of atoms.

Line	Column	Content
1	1	'ref' string specifying the reference energy
1	2	reference energy (in kcal/mol)
2	1	'molc' string specifying start of the first molecular block
2	2	unique integer identifier, 1 indicating the peptide conformer
2	3	charge of the peptide conformer
2	4	multiplicity of the peptide conformer
3,…,N+2	1	element type
3,…,N+2	2	X coordinates (in Å)
3,…,N+2	3	Y coordinates (in Å)
3,…,N+2	4	Z coordinates (in Å)
N+3	1	'end' string specifying end of the first molecular block
N+4	1	'molc' string specifying start of the second molecular block

Line	Column	Content
N+4	2	unique integer identifier, -1 indicating the lowest-energy conformer of the peptide
N+4	3	charge of the lowest-energy conformer of the peptide
N+4	4	multiplicity of the lowest-energy conformer of the peptide
N+4,…,2N+4	1	element type
N+4,…,2N+4	2	X coordinates (in Å)
N+4,…,2N+4	3	Y coordinates (in Å)
N+4,…,2N+4	4	Z coordinates (in Å)
2N+5	1	'end' string specifying end of the second molecular block

6. Technical Validation

The LC-ωPBE-XDM/aug-cc-pVTZ method was chosen as the reference level of theory for the single-point energy calculations of all the conformers in the PEPCONF data set. To justify the use of LC-ωPBE-XDM/aug-cc-pVTZ as the reference level, we checked its performance on several benchmark sets for conformational energies from the literature. The performance of LC-ωPBE-XDM/aug-cc-pVTZ is quantified in terms of the mean absolute error (MAE) relative to higher-level reference data. For Hobza's 2008 conformer database of small peptides[1], the MAE of LC-ωPBE-XDM/aug-cc-pVTZ relative to the CCSD(T)/CBS reference energies is 0.52 kcal/mol. The LC-ωPBE-XDM/aug-cc-pVTZ method also yields an MAE of 0.48 kcal/mol for the YMPJ[8] set of amino acid conformers relative to the MP2-F12/cc-pVTZ-F12 + [CCSD(Ts)-F12b – MP2-F12]/cc-pVDZ-F12 data. The MAE of LC-ωPBE-XDM/aug-cc-pVTZ for the smaller peptide conformer sets are as follows: 0.62 kcal/mol for CYCONF[4,11] (relative to CCSD(T)/CBS), 0.61 kcal/mol for PCONF[2] (relative to CCSD(T**)-F12a/CBS) and 0.60 kcal/mol for TPCONF[3,5] (relative to CCSD(T)/CBS).

Although they do not involve peptides, there are several other sets that can be used to validate the performance of LC-ωPBE-XDM/aug-cc-pVTZ for its ability to predict conformer energies. For example: 0.12 kcal/mol for ACONF[11,96] (*n*-alkane conformations, relative to W1h-val), 0.07 kcal/mol for BUT14DIOL[97] (conformations of butane-1,4-diol, relative to CCSD(T)-F12b/cc-pVTZ-F12), 0.75 kcal/mol for CCONF[98] (conformations of glucose and α-maltose, relative to DLPNO-CCSD(T)/CBS), 0.21 kcal/mol for MCONF[99] (melatonin conformations, relative to CCSD(T)/CBS), 0.24 kcal/mol for SCONF[11,100] (sugar conformations, relative to CCSD(T)/CBS), and 0.62 kcal/mol for UpU46[101] (RNA backbone conformations, relative to DLPNO-CCSD(T)/CBS). For comparison with peptide based non-covalent interaction energy data sets, LC-ωPBE-XDM/aug-cc-pVTZ gives MAE of 0.33 and 0.23 kcal/mol relative to DW-CCSD(T)-F12/aug-cc-pV(D+d)z for the BBI[102] and SSI[102] sets of backbone-backbone and sidechain-sidechain interactions, respectively. LC-ωPBE-XDM/aug-cc-pVTZ also yields an MAE of 0.28 and 0.18 kcal/mol for the S22 and S66 sets and 0.23 and 0.15 kcal/mol for the S22x5 and S66x8 sets of non-covalent binding energies calculated at the CCSD(T)/CBS limit, respectively.[103–107] A detailed analysis of the LC-ωPBE-XDM/aug-cc-pVTZ method for non-covalent interactions and thermochemistry can also be found in Ref. 108.

Data Citations

(1) Prasad, V. K., Otero-de-la-Roza, A. & DiLabio, G. A. *Figshare* http://dx.doi.org/10.6084/m9.figshare.7185194 (2018).

References

(1) Valdés, H., Pluháčková, K., Pitonák, M., Řezáč, J. & Hobza, P. Benchmark database on isolated small peptides containing an aromatic side chain: comparison between wave function and density functional theory methods and empirical force field. *Phys. Chem. Chem. Phys.* **10,** 2747-2757 (2008).

(2) Smith, D. G. A., Burns, L. A., Patkowski, K. & Sherrill, C. D. Revised damping parameters for the D3 dispersion correction to density functional theory. *J. Phys. Chem. Lett.* **7,** 2197–2203 (2016).

(3) Goerigk, L. *et al.* A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **19,** 32184–32215 (2017).

(4) Wilke, J. J., Lind, M. C., Schaefer III, H. F., Császár, A. G. & Allen, W. D. Conformers of gaseous cysteine. *J. Chem. Theory Comput.* **5,** 1511–1523 (2009).

(5) Goerigk, L., Karton, A., Martin, J. M. L. & Radom, L. Accurate quantum chemical energies for tetrapeptide conformations: why MP2 data with an insufficient basis set should be handled with caution. *Phys. Chem. Chem. Phys.* **15,** 7028-7031 (2013).

(6) Ropo, M., Schneider, M., Baldauf, C. & Blum, V. First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids. *Sci. Data* **3,** 160009 (2016).

(7) Yuan, Y., Mills, M. J. L., Popelier, P. L. A. & Jensen, F. Comprehensive analysis of energy minima of the 20 natural amino acids. *J. Phys. Chem. A* **118,** 7876–7891 (2014).

(8) Kesharwani, M. K., Karton, A. & Martin, J. M. L. Benchmark ab initio conformational energies for the proteinogenic amino acids through explicitly correlated methods. Assessment of density functional methods. *J. Chem. Theory Comput.* **12,** 444–454 (2016).

(9) Řezáč, J., Bím, D., Gutten, O. & Rulíšek, L. Toward accurate conformational energies of smaller peptides and medium-sized macrocycles: MPCONF196 benchmark energy data set. *J. Chem. Theory Comput.* **14,** 1254–1266 (2018).

(10) Goerigk, L. & Grimme, S. A general database for main group thermochemistry, kinetics, and noncovalent interactions─assessment of common and reparameterized (*meta*-)GGA density functionals. *J. Chem. Theory Comput.* **6,** 107–126 (2010).

(11) Goerigk, L. & Grimme, S. Efficient and accurate double-hybrid-meta-GGA density functionals − evaluation with the extended GMTKN30 database for general main group thermochemistry, kinetics, and noncovalent interactions. *J. Chem. Theory Comput.* **7,** 291–309 (2011).

(12) Sakae, Y. & Okamoto, Y. in *Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes: From Bioinformatics to Molecular Quantum Mechanics.* Springer Series in Bio-/Neuroinformatics Vol. 1 (ed. Liwo, A.) Ch. 7 (Springer, Berlin, Heidelberg, 2014).

(13) Lopes, P. E. M., Guvench, O. & MacKerell Jr., A. D. in *Molecular Modeling of Protien*s 2nd edn. Methods in Molecular Biology (Methods and Protocols) Vol. 1215 (ed. Kukol, A.) Ch. 3 (Humana Press, New York, NY, 2015).

(14) Huang, J. & MacKerell, A. D. Force field development and simulations of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **48,** 40–48 (2018).

(15) Dilabio, G. A. in *Non-Covalent Interactions in Quantum Chemistry and Physics: Theory and Applications* 1st edn (eds. Otero-de-la-Roza, A. & Dilabio, G. A.) Ch. 7 (Elsevier Inc., 2017).

(16) Prasad, V. K., Otero-de-la-Roza, A. & DiLabio, G. A. Atom-centered potentials with dispersion-corrected minimal-basis-set Hartree–Fock: an efficient and accurate computational approach for large molecular systems. *J. Chem. Theory Comput.* **14,** 726–738 (2018).

(17) Kruse, H. & Grimme, S. A geometrical correction for the inter- and intra-molecular basis set superposition error in

Hartree-Fock and density functional theory calculations for large systems. *J. Chem. Phys.* **136,** 154101 (2012).

(18) Řezáč, J. & Hobza, P. Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods. *J. Chem. Theory Comput.* **8,** 141–151 (2012).

(19) Witte, J., Neaton, J. B. & Head-Gordon, M. Effective empirical corrections for basis set superposition error in the def2-SVPD basis: gCP and DFT-C. *J. Chem. Phys.* **146,** 234105 (2017).

(20) Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinforma.* **78,** 1950-1958 (2010).

(21) Maier, J. A. *et al.* ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11,** 3696–3713 (2015).

(22) Wang, L.-P. *et al.* Building a more predictive protein force field: a systematic and reproducible route to AMBER-FB15. *J. Phys. Chem. B* **121,** 4023–4039 (2017).

(23) MacKerell Jr., A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102,** 3586–3616 (1998).

(24) Best, R. B. *et al.* Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles. *J. Chem. Theory Comput.* **8,** 3257–3273 (2012).

(25) Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14,** 71–73 (2017).

(26) Robertson, M. J., Tirado-Rives, J. & Jorgensen, W. L. Improved peptide and protein torsional energetics with the OPLS-AA force field. *J. Chem. Theory Comput.* **11,** 3499–3509 (2015).

(27) Shi, Y. *et al.* Polarizable atomic multipole-based AMOEBA force field for proteins. *J. Chem. Theory Comput.* **9,** 4046–4063 (2013).

(28) Thiel, W. Semiempirical quantum-chemical methods. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4,** 145–157 (2014).

(29) Brandenburg, J. G., Hochheim, M., Bredow, T. & Grimme, S. Low-cost quantum chemical methods for noncovalent interactions. *J. Phys. Chem. Lett.* **5,** 4275–4284 (2014).

(30) Christensen, A. S., Kubař, T., Cui, Q. & Elstner, M. Semiempirical quantum mechanical methods for noncovalent interactions for chemical and biochemical applications. *Chem. Rev.* **116,** 5301–5337 (2016).

(31) Dewar, M. J. S., Zoebisch, E. G., Healy, E. F. & Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **107,** 3902–3909 (1985).

(32) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **13,** 1173–1213 (2007).

(33) Řezáč, J., Fanfrlík, J., Salahub, D. & Hobza, P. Semiempirical quantum chemical PM6 method augmented by dispersion and H-bonding correction terms reliably describes various types of noncovalent complexes. *J. Chem. Theory Comput.* **5,** 1749–1760 (2009).

(34) Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **19,** 1–32 (2013).

(35) Tuttle, T. & Thiel, W. OMx-D: semiempirical methods with orthogonalization and dispersion corrections. Implementation and biochemical application. *Phys. Chem. Chem. Phys.* **10,** 2159-2166 (2008).

(36) Dral, P. O. *et al.* Semiempirical quantum-chemical orthogonalization-corrected methods: theory, implementation, and parameters. *J. Chem. Theory Comput.* **12,** 1082–1096 (2016).

(37) Frauenheim, Th. *et al.* A self-consistent charge density-functional based tight-binding method for predictive materials simulations in physics, chemistry and biology. *physica status solidi (b)* **217,** 41–62 (2000).

(38) Koskinen, P. & Mäkinen, V. Density-functional tight-binding for beginners. *Comput. Mater. Sci.* **47,** 237–253 (2009).

(39) Krishnapriyan, A., Yang, P., Niklasson, A. M. N. & Cawkwell, M. J. Numerical optimization of density functional tight binding models: application to molecules containing carbon, hydrogen, nitrogen, and oxygen. *J. Chem. Theory Comput.* **13,** 6191–6200 (2017).

(40) Grimme, S., Bannwarth, C. & Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (*Z*=1–86). *J. Chem. Theory Comput.* **13,** 1989–2009 (2017).

(41) Sure, R. & Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **34,** 1672–1685 (2013).

(42) Goerigk, L. & Reimers, J. R. Efficient methods for the quantum chemical treatment of protein structures: the effects of

London-dispersion and basis-set incompleteness on peptide and water-cluster geometries. *J. Chem. Theory Comput.* **9,** 3240–3251 (2013).

(43) Goerigk, L., Collyer, C. A. & Reimers, J. R. Recommending Hartree–Fock theory with London-dispersion and basis-set-superposition corrections for the optimization or quantum refinement of protein structures. *J. Phys. Chem. B* **118,** 14612–14626 (2014).

(44) Grimme, S., Brandenburg, J. G., Bannwarth, C. & Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *J. Chem. Phys.* **143,** 054107 (2015).

(45) Sure, R., Brandenburg, J. G. & Grimme, S. Small atomic orbital basis set first-principles quantum chemical methods for large molecular and periodic systems: a critical analysis of error sources. *ChemistryOpen* **5,** 94–109 (2016).

(46) Brandenburg, J. G., Caldeweyher, E. & Grimme, S. Screened exchange hybrid density functional for accurate and efficient structures and interaction energies. *Phys. Chem. Chem. Phys.* **18,** 15519–15523 (2016).

(47) Brandenburg, J. G., Bannwarth, C., Hansen, A. & Grimme, S. B97-3c: A revised low-cost variant of the B97-D density functional method. *J. Chem. Phys.* **148,** 064104 (2018).

(48) Hamley, I. W. Small bioactive peptides for biomaterials design and therapeutics. *Chem. Rev.* **117,** 14015–14041 (2017).

(49) Case, D. A. *et al. AMBER 2016* (University of California, 2016).

(50) Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26,** 1668–1688 (2005).

(51) Salomon-Ferrer, R., Case, D. A. & Walker, R. C. An overview of the Amber biomolecular simulation package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **3,** 198–210 (2013).

(52) RCSB Protein Data Bank. https://www.rcsb.org/ (2018).

(53) Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

(54) The Cambridge Structural Database. https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/ (2018).

(55) Groom, C. R., Bruno, I. J., Lightfoot, M. P., & Ward, S. C. The cambridge structural database. *Acta Cryst. B,* **72,** 171–179 (2016).

(56) The Open Babel Package, version 2.3.2. http://openbabel.org (2018).

(57) O'Boyle, N. M. *et al.* Open Babel: an open chemical toolbox. *J. Cheminform.* **3,** 33 (2011).

(58) Avogadro: An Advanced Molecule Editor and Visualizer. https://avogadro.cc/ (2018).

(59) Hanwell, M. D. *et al.* Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminform.* **4,** 17 (2012).

(60) Bruccoleri, R. E. & Karplus, M. Conformational sampling using high-temperature molecular dynamics. *Biopolymers* **29,** 1847–1862 (1990).

(61) Settanni, G. & Fersht, A. R. High temperature unfolding simulations of the TRPZ1 peptide. *Biophys. J.* **94,** 4444–4453 (2008).

(62) Walczewska-Szewc, K., Deplazes, E. & Corry, B. Comparing the ability of enhanced sampling molecular dynamics methods to reproduce the behavior of fluorescent labels on proteins. *J. Chem. Theory Comput.* **11,** 3455–3465 (2015).

(63) Dalby, A. & Shamsir, M. S. Molecular Dynamics Simulations of the Temperature Induced Unfolding of Crambin Follow the Arrhenius Equation. *F1000Research* **4,** 589 (2015).

(64) Neale, C., Pomès, R. & García, A. E. Peptide bond isomerization in high-temperature simulations. *J. Chem. Theory Comput.* **12,** 1989–1999 (2016).

(65) RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org/ (2018).

(66) Ebejer, J.-P., Morris, G. M. & Deane, C. M. Freely available conformer generation methods: how good are they? *J. Chem. Inf. Model.* **52,** 1146–1158 (2012).

(67) Blaney, J. M. & Dixon, J. S. in *Reviews in Computational Chemistry* Vol. 5 (eds. Lipkowitz, K. B. & Boyd, D. B.) Ch. 6 (John Wiley & Sons, Inc., 2007).

(68) Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **4,** 170193 (2017).

(69) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38,** 3098–3100 (1988).

(70) Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37,** 785–789 (1988).

(71) Hariharan, P. C. & Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor. Chim. Acta* **28,** 213–222 (1973).

(72) Francl, M. M. *et al.* Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row

elements. *J. Chem. Phys.* **77,** 3654–3665 (1982).

(73) Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132,** 154104 (2010).

(74) Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32,** 1456–1465 (2011).

(75) Becke, A. D. & Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction. *J. Chem. Phys.* **122,** 154104 (2005).

(76) Johnson, E. R. & Becke, A. D. A post-Hartree–Fock model of intermolecular interactions. *J. Chem. Phys.* **123,** 024101 (2005).

(77) Becke, A. D. & Johnson, E. R. A density-functional model of the dispersion interaction. *J. Chem. Phys.* **123,** 154101 (2005).

(78) Becke, A. D. & Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction: high-order dispersion coefficients. *J. Chem. Phys.* **124,** 014104 (2006).

(79) Johnson, E. R. & Becke, A. D. A post-Hartree-Fock model of intermolecular interactions: inclusion of higher-order corrections. *J. Chem. Phys.* **124,** 174104 (2006).

(80) Becke, A. D. & Johnson, E. R. A unified density-functional treatment of dynamical, nondynamical, and dispersion correlations. *J. Chem. Phys.* **127,** 124108 (2007).

(81) Becke, A. D. & Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction revisited. *J. Chem. Phys.* **127,** 154108 (2007).

(82) Otero-de-la-Roza, A. & DiLabio, G. A. Transferable atom-centered potentials for the correction of basis set incompleteness errors in density-functional theory. *J. Chem. Theory Comput.* **13,** 3505–3524 (2017).

(83) Frisch, M. J. *et al. Gaussian 09, Revision D.01* (Gaussian, Inc., 2009).

(84) Frisch, M. J. *et al. Gaussian 16, Revision B.01* (Gaussian, Inc., 2016).

(85) Huzinaga, S. *et al.* in *Gaussian Basis Sets for Molecular Calculations.* Physical Sciences Data Series Vol. 16 1st edn (Elsevier Inc., 1984).

(86) Vydrov, O. A. & Scuseria, G. E. Assessment of a long-range corrected hybrid functional. *J. Chem. Phys.* **125,** 234109 (2006).

(87) Vydrov, O. A., Heyd, J., Krukau, A. V. & Scuseria, G. E. Importance of short-range versus long-range Hartree-Fock exchange for the performance of hybrid density functionals. *J. Chem. Phys.* **125,** 074106 (2006).

(88) Dunning Jr., T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **90,** 1007–1023 (1989).

(89) Kendall, R. A., Dunning Jr., T. H. & Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **96,** 6796–6806 (1992).

(90) Woon, D. E. & Dunning Jr., T. H. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *J. Chem. Phys.* **98,** 1358–1371 (1993).

(91) Otero-de-la-Roza, A., Johnson, E. R. & DiLabio, G. A. Halogen bonding from dispersion-corrected density-functional theory: the role of delocalization error. *J. Chem. Theory Comput.* **10,** 5436–5447 (2014).

(92) Hohenstein, E. G. & Sherrill, C. D. Wavefunction methods for noncovalent interactions. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2,** 304–326 (2012).

(93) Řezáč, J. & Hobza, P. Describing noncovalent interactions beyond the common approximations: how accurate is the "Gold Standard," CCSD(T) at the complete basis set limit? *J. Chem. Theory Comput.* **9,** 2151–2155 (2013).

(94) Riplinger, C. & Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.* **138,** 034106 (2013).

(95) Riplinger, C., Sandhoefer, B., Hansen, A. & Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *J. Chem. Phys.* **139,** 134101 (2013).

(96) Gruzman, D., Karton, A. & Martin, J. M. L. Performance of ab initio and density functional methods for conformational equilibria of C_nH_{2n+2} alkane isomers (n=4−8). *J. Phys. Chem. A* **113,** 11974–11983 (2009).

(97) Kozuch, S., Bachrach, S. M. & Martin, J. M. L. Conformational equilibria in butane-1,4-diol: a benchmark of a prototypical system with strong intramolecular H-bonds. *J. Phys. Chem. A* **118,** 293–303 (2014).

(98) Marianski, M., Supady, A., Ingram, T., Schneider, M. & Baldauf, C. Assessing the accuracy of across-the-scale methods for predicting carbohydrate conformational energies for the examples of glucose and α-maltose. *J. Chem. Theory Comput.* **12,** 6157–6168 (2016).

(99) Fogueri, U. R., Kozuch, S., Karton, A. & Martin, J. M. L. The melatonin conformer space: benchmark and assessment of wave function and DFT methods for a paradigmatic biological and pharmacological molecule. *J. Phys. Chem. A* **117,** 2269–2277 (2013).

(100) Csonka, G. I., French, A. D., Johnson, G. P. & Stortz, C. A. Evaluation of density functionals and basis sets for carbohydrates. *J. Chem. Theory Comput.* **5,** 679–692 (2009).

(101) Kruse, H. *et al.* Quantum chemical benchmark study on 46 RNA backbone families using a dinucleotide unit. *J. Chem. Theory Comput.* **11,** 4972–4991 (2015).

(102) Burns, L. A. *et al.* The biofragment database (BFDb): an open-data platform for computational chemistry analysis of noncovalent interactions. *J. Chem. Phys.* **147,** 161727 (2017).

(103) Jurečka, P., Šponer, J., Černý, J. & Hobza, P. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.* **8,** 1985–1993 (2006).

(104) Gráfová, L., Pitoňák, M., Řezáč, J. & Hobza, P. Comparative study of selected wave function and density functional methods for noncovalent interaction energy calculations using the extended S22 data set. *J. Chem. Theory Comput.* **6,** 2365–2376 (2010).

(105) Řezáč, J., Riley, K. E. & Hobza, P. S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures. *J. Chem. Theory Comput.* **7,** 2427–2438 (2011).

(106) Řezáč, J., Riley, K. E. & Hobza, P. Extensions of the S66 data set: more accurate interaction energies and angular-displaced nonequilibrium geometries. *J. Chem. Theory Comput.* **7,** 3466–3470 (2011).

(107) Brauer, B., Kesharwani, M. K., Kozuch, S. & Martin, J. M. L. The S66x8 benchmark for noncovalent interactions revisited: explicitly correlated ab initio methods and density functional theory. *Phys. Chem. Chem. Phys.* **18,** 20905–20925 (2016).

(108) Otero-de-la-Roza, A. & Johnson, E. R. Non-covalent interactions and thermochemistry using XDM-corrected hybrid and range-separated hybrid density functionals. *J. Chem. Phys.* **138,** 204109 (2013).

Chapter 5

BSE49, a diverse, high-quality benchmark dataset of separation energies of chemical bonds

Abstract

We present an extensive and diverse dataset of bond separation energies associated with the homolytic cleavage of covalently bonded molecules (A-B) into their corresponding radical fragments ($A^{\cdot}$ and $B^{\cdot}$). Our dataset contains two different classifications of model structures referred to as "*Existing"* (molecules with associated experimental data) and *"Hypothetical"* (molecules with no associated experimental data). In total, the dataset consists of 4502 datapoints (1969 datapoints from the *Existing* and 2533 datapoints from the *Hypothetical* classes). The dataset covers 49 unique X-Y type single bonds (except H-H, H-F, and H-Cl), where X and Y are H, B, C, N, O, F, Si, P, S, and Cl atoms. All the reference data was calculated at the (RO)CBS-QB3 level of theory. The reference bond separation energies are non-relativistic ground-state energy differences and contain no zero-point energy corrections. This new dataset of bond separation energies (BSE49) is presented as a high-quality reference dataset for assessing and developing computational chemistry methods.

1. Background & Summary

Bond dissociation enthalpies (BDEs) are a central property in chemistry that have been studied for decades experimentally and computationally.[1–4] BDEs can be used to estimate the selectivity and reactivity of various molecules with free radicals (like $^{\cdot}OH$, $^{\cdot}OOH$, $^{\cdot}OR$, $^{\cdot}OOR$, $^{\cdot}NO$, $^{\cdot}NO_2$, etc.) that are generated and transformed during chemical reactions relevant in chemistry and biology.[5–10] In this context, the calculation of BDEs for C-H, O-H, N-H, S-H, O-O, and S-S bonds in biologically relevant systems can help develop an understanding of the efficiency of antioxidants.[11–13] Furthermore, the calculation of BDEs is fundamental to develop a deeper understanding of various enzyme catalytic processes[14–16] and surface functionalization chemistry[17–19].

In 2012, Drew and Reynisson employed BDE calculations to predict the major metabolic sites of fifty known drug molecules.[20] Similarly, Andersson and co-workers applied BDE calculations to estimate the sensitivity of various drug candidates toward autoxidation.[21] The application of computed BDEs in these works shows how computational techniques can be incorporated into the risk assessment of drug products and guide further experimentation. Computationally obtained BDEs were also reported in different studies[22–24], where the C-O and C-C BDEs were calculated for several substituted analogues of lignin, an abundant polymeric organic material and a potential renewable source of biofuels and chemicals.[22–24] The calculated BDEs were used to predict the homolytic dissociation of C-C and C-O bonds under thermal decomposition using model compounds representing the dominant linkages of lignin.

Given the importance of BDEs in many areas of chemistry and, consequently, the need to accurately predict bond energies computationally, a dataset of accurately predicted bond separation energies (BSEs) is developed here using an accurate computational chemistry method. Bond separation energies are a molecular property that can be computed in a straightforward manner in vacuum and provides direct information about the strength of a chemical bond. The BSEs presented in this work are differences between non-relativistic ground-state energies and contain no vibrational energy contributions, no zero-point energies, and no attempt has been made at thermally averaging over molecular conformations. As such, the reported BSEs are not comparable to experimental BDEs, but they serve as an ideal resource for developing and evaluating lower-cost computational chemistry methods used for a wide range of applications in chemistry and biology. Similar datasets to the one proposed in this work are available in the literature, but they tend to be small in terms of the total number of datapoints[25], lack bond-type diversity[26,27] or are calculated using less accurate computational chemistry methods compared to the one used in this work[28–30]. To the best of our knowledge, an accurate and extensive dataset of computationally predicted BSEs is not available in the literature. The main reason for this absence is that BSE calculations with high accuracy require computationally expensive methods that tend to scale poorly with system size.

This work addresses the aforementioned gap in the literature by constructing a large dataset (4502 datapoints) of computationally predicted BSEs of 49 unique bond types, all of which are determined with a high-level composite theoretical procedure denoted as (RO)CBS-QB3[31–33]. This approach ensures uniform, high-quality reference data and eliminates the need to collect and verify data gathered from various sources, which may differ substantially in their accuracy. The (RO)CBS-QB3 method is known to produce BDEs of high accuracy.[8,33–37] Therefore, it is suitable for developing a database of BSEs that can be used to test and parametrize low-cost computational methods. One particular target application of our dataset is for the training of cost-effective computational approaches like atom-centered potentials[38–40] (ACP) or machine learning potentials[28–30].

2. Methods

Dataset composition

We present the BSE49 dataset, which comprises a broad range of bond separation energies for 49 unique bond types. The model systems present in the dataset are neutral molecules with X-H, X-F, X-Cl, X-X, and X-Y single bonds, where X and Y are B, C, N, O, Si, P, and S. The number of datapoints and the ranges of bond separation energies associated with each bond type are provided in Table 1. The structures of model systems on which the calculations were performed are divided into *"Existing"* and "*Hypothetical"* classes. The *Existing* type structures were built by selecting molecules with experimental data reported in the *Comprehensive Handbook of Chemical Bond Dissociation Energies*[41]. In contrast, the *Hypothetical* type structures were constructed by functional group substitutions of X-Y single bonds in order to include bond types that were not present in the handbook and to increase the diversity and number of datapoints for each bond type in the dataset. The candidate molecules for both *Existing* and *Hypothetical* subsets were generated using a partially automated computational workflow as described below.

Table 1. List of the number of datapoints in the BSE49 dataset and the ranges of bond separation energies associated with each bond type calculated using (RO)CBS-QB3.[a]

Bond type	Data points	Range of bond separation energies
B-H	68	77.22 - 115.14
C-H	395	80.08 - 141.22
N-H	156	53.05 - 131.63
O-H	240	68.65 - 126.75
Si-H	111	74.31 - 106.06
P-H	118	61.73 - 87.98
S-H	39	74.80 - 95.81
B-B	75	47.41 - 112.40
B-C	83	92.26 - 142.78
B-N	71	85.50 - 155.16
B-O	51	100.14 - 158.50
B-F	82	152.61 - 177.24
B-Si	84	36.27 - 110.83
B-P	89	72.64 - 99.12
B-S	51	84.10 - 128.28
B-Cl	81	81.86 - 128.98
C-C	363	64.69 - 156.08
C-N	98	27.65 - 122.95
C-O	171	48.31 - 127.45
C-F	40	103.44 - 133.45

Bond type	Data points	Range of bond separation energies
C-Si	153	36.82 - 111.67
C-P	85	60.93 - 115.15
C-S	64	41.42 - 105.29
C-Cl	129	64.26 - 113.54
N-N	37	15.64 - 70.81
N-O	31	22.50 - 70.80
N-F	36	49.72 - 83.45
N-Si	64	33.93 - 122.94
N-P	93	40.82 - 91.06
N-S	53	24.53 - 72.61
N-Cl	31	35.89 - 80.63
O-O	60	21.20 - 56.42
O-F	90	11.04 - 51.79
O-Si	144	74.85 - 144.88
O-P	27	83.10 - 130.79
O-S	51	46.55 - 93.05
O-Cl	85	9.38 - 61.56
F-Si	36	123.92 - 169.04
F-P	32	99.43 - 125.94
F-S	99	72.84 - 107.41
Si-Si	165	34.86 - 104.94
Si-P	65	60.09 - 87.04
Si-S	57	62.95 - 98.18
Si-Cl	102	109.68 - 123.12
P-P	20	44.37 - 77.37
P-S	29	67.42 - 96.01
P-Cl	32	69.91 - 89.30
S-S	64	37.09 - 78.33
S-Cl	102	50.44 - 71.17

[a] the bond separation energy ranges are in kcal/mol.

Dataset generation

The calculated bond separation energies are defined as the negative of the difference in the ground-state electronic energies for the reaction

$$A\text{-}B \rightarrow A^{\cdot} + B^{\cdot}$$

where $A^{\cdot}$ and $B^{\cdot}$ represent the two radical fragments formed by homolytically breaking the A-B covalent bond in vacuum. Based on this reaction, the equilibrium geometries of the parent molecules and their respective radical fragments are required to calculate the bond separation energies. The geometries of the parent molecule and the associated radicals were constructed manually for both *Existing* and *Hypothetical* subsets using the *Avogadro*[42] program. The constructed geometries were then used as starting points for a conformer search. The *CSD conformer generator*[43] and *FullMonte*[44] codes were used to generate multiple conformers. The geometry of each conformer was relaxed to the corresponding local minimum using the *Gaussian*[45] software package. This relaxation was carried out first by using a low-level method, combining the B3LYP[46–51] density functional and 6-31G*[52,53] basis set along with the D3[54–56] dispersion correction scheme using the Becke-Johnson[57] damping (B3LYP-D3(BJ)/6-31G*). The optimized conformers were ranked using the B3LYP-D3(BJ)/6-31G* relative energies at the local minima. The ten lowest-energy conformers were then re-optimized at the higher-level CAM-B3LYP-D3(BJ)/def2-TZVP level of theory[54–59]. Range-separated functionals like CAM-B3LYP minimize the delocalization error, which could be important in the description of radical species.[60] The lowest-energy conformer obtained in this procedure was used for calculating the bond separation energies using the composite method described below. All calculations employed a default self-consistent field (SCF) convergence criterion of 10^{-8} Hartrees, *ultrafine* integration grid, and a *tight* optimization convergence criteria (maximum force = 1.5 x 10^{-5} Hartrees/Bohr, RMS force = 1 x 10^{-5} Hartrees/Bohr, maximum displacement = 6 x 10^{-5} Bohr, RMS displacement = 4 x 10^{-5} Bohr).

This partially automated workflow produced structures that are not necessarily the global minima. A visual inspection of the structures revealed that about 20% of the conformers generated do not correspond to the global minima, which reflects the difficulty of solving a global optimization problem (finding the most stable conformer) for such a large number of systems reliably. In addition, due to computational constraints, no attempt was made at evaluating the conformational energy landscape and statistically weighting the low-energy conformers associated with each molecule. Therefore, the dataset is not appropriate for direct comparison to bond separation energies obtained by back-correcting experimental BDEs, but it is suitable for testing and training computationally less expensive methods regarding their ability to accurately calculate the energy difference between the chosen conformers of products (A· and B·) and reactant (A-B).

The structures obtained from the workflow described above were then used for the final step of reference data calculation, using the composite (RO)CBS-QB3[31–33] method. The restricted-open-shell[61] CBS-QB3 or ROCBS-QB3 was employed for the open-shell radical fragments, while restricted closed-shell calculations were performed for the closed-shell parent molecules with CBS-QB3. The composite (RO)CBS-QB3 method approximates energies at the complete-basis-set CCSD(T) level, using a series of computationally lower-cost methods including: (i) geometry optimization followed by vibration frequency calculation using the unrestricted-open-shell[62] B3LYP/6-311G(2d,d,p) method[46–51,63], (ii) ROMP2/6-

311+G(3d2f,2df,2p) level[63–65] energy extrapolated to the complete-basis-set limit, (iii) energy calculation at ROMP4(SDQ)/6-31+G(d(f),p) level[63,64,66], and (iv) energy calculation at ROCCSD(T)/6-31+G† level[63,64,67] (where 6-31+G† is a modified 6-31+G(d) basis set). Note that the final (RO)CBS-QB3 energy includes additional empirical correction terms described in Reference 33. Structures were screened to remove any system for which the imaginary frequencies were obtained. The (RO)CBS-QB3 energies for the structures associated with a particular bond breaking reaction were used to obtain the bond separation energies for the dataset.

3. Data Records

The reference bond separation energies (in kcal/mol) and coordinates (in Å) of the structures presented in the BSE49 dataset are publicly available free-of-charge from the Figshare[68] and GitHub (https://github.com/aoterodelaroza/bse49) repositories in the plain-text database file format (DB format) described in Table 2. The atomic coordinates of the model structures are stored in a plain-text XYZ format in the *Geometries* directory. The BSE49 dataset contains one DB format file and three XYZ format files for each bond separation energy. In total, deposited files include 4502 DB format files stored in the *db-BSE49* directory and 13506 XYZ format files stored in their respective *Existing* or *Hypothetical* classification directories. Additional files labelled as *BSE49_Existing.org* and *BSE49_Hypothetical.org* are also provided. These files contain the necessary information about the reference data for all the model systems.

4. File Format

For each molecule, the reference bond separation energy and the atomic coordinates are stored in a file named *MoleculeName.db*. The Cartesian coordinates of the atoms are stored in files called *MoleculeName_AB.xyz, MoleculeName_A.xyz*, and *MoleculeName_B.xyz*, where *AB* represents the parent molecule, *A* represents the first radical fragment, and *B* represents the second radical fragment.

The DB format file contains a header line specifying the reference energy value (in kcal/mol) followed by three '*molc'* (short for molecule) blocks containing a unique integer identifier, charge, multiplicity, and the atomic coordinates (in Å) of the parent molecule and its corresponding radical fragments. The XYZ format file contains a header line defining the number of atoms N, a comment line containing the charge and multiplicity, and N lines with each containing element type and X, Y, Z coordinates (in Å). The *BSE49_Existing.org* and *BSE49_Hypothetical.org* files are special-character separated plain-text files (where the special character is '|') containing multiple lines and eight columns.

The columns are: (i) dataset name of the model system, (ii) unique integer identifier 1 indicating the A· fragment, (iii) geometry filename of the A· fragment, (iv) unique integer identifier 1 indicating the B· fragment, (v) geometry filename of the B· fragment, (vi) unique integer identifier -1 indicating the A-B model system, (vii) geometry filename of the A-B model system, and (viii) computational reference bond separation energy (in kcal/mol).

Table 2. A description of the DB format file (.db) for an A-B molecule containing N number of atoms with two radical fragments (A· and B·), which have n_1 and n_2 number of atoms, respectively.

Line	Column	Content
1	1	'ref' string specifying reference energy
1	2	reference bond separation energy (in kcal/mol)
2	1	'molc' string specifying start of the first molecular block
2	2	unique integer identifier, 1 indicating the A· fragment
2	3	the charge of the A· fragment
2	4	the multiplicity of the A· fragment
$3,\ldots,n_1+2$	1	element type
$3,\ldots,n_1+2$	2	X coordinates (in Å)
$3,\ldots,n_1+2$	3	Y coordinates (in Å)
$3,\ldots,n_1+2$	4	Z coordinates (in Å)
n_1+3	1	'end' string specifying end of the first molecular block
n_1+4	1	'molc' string specifying start of the second molecular block
n_1+4	2	unique integer identifier, 1 indicating B· fragment
n_1+4	3	the charge of the B· fragment
n_1+4	4	the multiplicity of the B· fragment
$n_1+5,\ldots,n_1+n_2+4$	1	element type
$n_1+5,\ldots,n_1+n_2+4$	2	X coordinates (in Å)
$n_1+5,\ldots,n_1+n_2+4$	3	Y coordinates (in Å)
$n_1+5,\ldots,n_1+n_2+4$	4	Z coordinates (in Å)
n_1+n_2+5	1	'end' string specifying end of the second molecular block
n_1+n_2+6	1	'molc' string specifying start of the third molecular block
n_1+n_2+6	2	unique integer identifier, -1 indicating the A-B parent molecule
n_1+n_2+6	3	the charge of the A-B parent molecule
n_1+n_2+6	4	the multiplicity of the A-B parent molecule
$n_1+n_2+7,\ldots,n_1+n_2+N+6$	1	element type
$n_1+n_2+7\ldots,n_1+ n_2+N+6$	2	X coordinates (in Å)
$n_1+n_2+7,\ldots,n_1+n_2+N+6$	3	Y coordinates (in Å)
$n_1+n_2+7,\ldots,n_1+n_2+N+6$	4	Z coordinates (in Å)
n_1+n_2+N+7	1	'end' string specifying end of the third molecular block

5. Technical Validation

For the generation of reference data, the reliable (RO)CBS-QB3 method was chosen for all the model systems considered in the BSE49 dataset. The (RO)CBS-QB3 method has been widely used in literature in recent years.[69–90] The developers of the (RO)CBS-QB3 method reported that it predicts heats of formation at 298K with a mean absolute deviation (MAD) from the experiment of 0.91 kcal/mol.[33] For bond dissociation enthalpies of eleven molecules with chemical structures typically found in amino acid sidechains, peptide termini, and peptide backbones, Moore *et al.* reported an MAD of 1.72 kcal/mol from the experimental values.[8] For small lignin model molecules, the CBS-QB3 approach was shown to yield bond dissociation enthalpies within 2.99 kcal/mol from experimental values.[34] (RO)CBS-QB3 has been used as a reference method for benchmarking various density functional theory methods to estimate bond dissociation enthalpies in a different study on small lignin model systems.[23] Hudzik and co-workers utilized the CBS-QB3 composite method to study the C-H bond separation energies of a few alkane molecules and reported a good agreement with literature values.[35] The (RO)CBS-QB3 has also been used for the prediction of bond dissociation enthalpies in a previous work by Menon *et al*.[36] The MAD of (RO)CBS-QB3 was reported to be only 0.60 kcal/mol from the experiment and was suggested as being a reliable and efficient procedure for calculating bond separation energies in comparison to the other composite methods tested. In another work, bond dissociation enthalpies of 200 molecules were calculated using an earlier version of this work's composite method, CBS-Q.[37] It was shown that the results of the CBS-Q composite procedure predicted bond dissociation enthalpies to within 2.39 kcal/mol of the reported experimental values. Collectively, these results support the selection of (RO)CBS-QB3 as a practical and accurate method for the generation of reference data in this work. Note that the reference bond separation energies reported in this work are non-relativistic (RO)CBS-QB3 energies without zero-point energy corrections. This makes the reference data suitable to support the development of low-cost computational chemistry methods like those described in references 28–30 and 38–40.

6. Code Availability

Throughout this work, the *Gaussian* software package was used for geometry optimizations, frequency calculations, and composite (RO)CBS-QB3 calculations. The *Gaussian* software package can be purchased from Gaussian Inc. (http://gaussian.com/) under a commercial license. *CSD conformer generator* was used for conformer generation. The *CSD conformer generator* can be purchased under a commercial license from https://www.ccdc.cam.ac.uk/solutions/csd-enterprise/applications/conformer-generator/. *Fullmonte* software package was also used along with *MOPAC16* (PM6-DH2 method).

Fullmonte software package can be downloaded free-of-cost from https://github.com/bobbypaton/FullMonte. Whereas *MOPAC16* software package can be installed after acquiring a free license from http://openmopac.net/. The *Avogadro* molecular editor and visualizer is an open-source program available at https://avogadro.cc/.

References

(1) Chan, B., Collins, E. & Raghavachari, K. Applications of isodesmic-type reactions for computational thermochemistry. *WIREs Comput. Mol. Sci.* **11**, e1501 (2020).

(2) Johnson, E. R., Clarkin, O. J. & DiLabio, G. A. Density functional theory based model calculations for accurate bond dissociation enthalpies. 3. A single approach for X-H, X-X, and X-Y (X, Y = C, N, O, S, Halogen) bonds. *J. Phys. Chem. A* **107**, 9953–9963 (2003).

(3) DiLabio, G. A. & Pratt, D. A. Density functional theory based model calculations for accurate bond dissociation enthalpies. 2. Studies of X-X and X-Y (X, Y = C, N, O, S, Halogen) bonds. *J. Phys. Chem. A* **104**, 1938–1943 (2000).

(4) DiLabio, G. A., Pratt, D. A., LoFaro, A. D. & Wright, J. S. Theoretical study of X-H bond energetics (X = C, N, O, S): Application to substituent effects, gas phase acidities, and redox potentials. *J. Phys. Chem. A* **103**, 1653–1661 (1999).

(5) Hioe, J. & Zipse, H. Radical stability–Thermochemical aspects. In *Encyclopedia of Radicals in Chemistry, Biology and Materials*, eds. Chatgilialoglu, C. & Studer, A. (John Wiley & Sons, Ltd., 2012).

(6) Zavitsas, A. A. Thermochemistry and hydrogen transfer kinetics. In *Encyclopedia of Radicals in Chemistry, Biology and Materials*, eds. Chatgilialoglu, C. & Studer, A. (John Wiley & Sons, Ltd., 2012).

(7) Hioe, J., Mosch, M., Smith, D. M. & Zipse, H. Dissociation energies of Cα–H bonds in amino acids –a re-examination. *RSC Adv.* **3**, 12403-12408 (2013).

(8) Moore, B. N. & Julian, R. R. Dissociation energies of X–H bonds in amino acids. *Phys. Chem. Chem. Phys.* **14**, 3148-3154 (2012).

(9) Coote, M. L. & Zavitsas, A. A. Using inherent radical stabilization energies to predict unknown enthalpies of formation and associated bond dissociation energies of complex molecules. *Tetrahedron* **72**, 7749–7756 (2016).

(10) Adhikary, A., Kumar, A., Becker, D. & Sevilla, M. D. Understanding DNA radicals employing theory and electron spin resonance spectroscopy. In *Encyclopedia of Radicals in Chemistry, Biology and Materials*, eds. Chatgilialoglu, C. & Studer, A. (John Wiley & Sons, Ltd., 2012).

(11) Wright, J. S., Johnson, E. R. & DiLabio, G. A. Predicting the activity of phenolic antioxidants: Theoretical method, analysis of substituent effects, and application to major families of antioxidants. *J. Am. Chem. Soc.* **123**, 1173-1183 (2001).

(12) Kancheva, V. D. *et al.* Antiradical and antioxidant activities of new bio-antioxidants. *Biochimie* **94**, 403–415 (2012).

(13) Bond dissociation energies and thermodynamic functions of antioxidants. In *Handbook of Antioxidants: Bond Dissociation Energies, Rate Constants, Activation Energies, and Enthalpies of Reactions, Second Edition*, eds. Denisov, E. T. & Denisova, T. (CRC Press LLC, 1999)

(14) Jensen, K. P. & Ryde, U. How the Co-C bond is cleaved in coenzyme B_{12} enzymes: A theoretical study. *J. Am. Chem. Soc.* **127**, 9117–9128 (2005).

(15) Qu, Z. W., Hansen, A. & Grimme, S. Co-C bond dissociation energies in cobalamin derivatives and dispersion effects: Anomaly or just challenging? *J. Chem. Theory Comput.* **11**, 1037–1045 (2015).

(16) Kozlowski, P. M. *et al.* The cobalt-methyl bond dissociation in methylcobalamin: New benchmark analysis based on density functional theory and completely renormalized coupled-cluster calculations. *J. Chem. Theory Comput.* **8**, 1870–1894 (2012).

(17) Kruse, P., Johnson, E. R., DiLabio, G. A. & Wolkow, R. A. Patterning of vinylferrocene on H-Si(100) via self-directed growth of molecular lines and STM-induced decomposition. *Nano Lett.* **2**, 807–810 (2002).

(18) Tong, X., DiLabio, G. A. & Wolkow, R. A. A self-directed growth process for creating covalently bonded molecular assemblies on the H-Si(100)-3x1 surface. *Nano Lett.* **4**, 979–983 (2004).

(19) Piva, P. G. *et al.* Field regulation of single-molecule conductivity by a charged surface atom. *Nature* **435**, 658–661

(2005).

(20) Drew, K. L. M. & Reynisson, J. The impact of carbon–hydrogen bond dissociation energies on the prediction of the cytochrome P450 mediated major metabolic site of drug-like compounds. *Eur. J. Med. Chem.* **56**, 48–55 (2012).

(21) Andersson, T., Broo, A. & Evertsson, E. Prediction of drug candidates' sensitivity toward autoxidation: Computational estimation of C-H dissociation energies of carbon-centered radicals. *J. Pharm. Sci.* **103**, 1949–1955 (2014).

(22) Parthasarathi, R., Romero, R. A., Redondo, A. & Gnanakaran, S. Theoretical study of the remarkably diverse linkages in lignin. *J. Phys. Chem. Lett.* **2**, 2660–2666 (2011).

(23) Kim, S. *et al.* Computational study of bond dissociation enthalpies for a large range of native and modified lignins. *J. Phys. Chem. Lett.* **2**, 2846–2852 (2011).

(24) Beste, A. & Buchanan III, A. C. Computational study of bond dissociation enthalpies for lignin model compounds. Substituent effects in phenethyl phenyl ethers. *J. Org. Chem.* **74**, 2837–2841 (2009).

(25) Chan, B. & Radom, L. BDE261: A comprehensive set of high-level theoretical bond dissociation enthalpies. *J. Phys. Chem. A* **116**, 4975–4986 (2012).

(26) Saito, T., Kambara, H. & Takano, Y. Quantitative assessment of reparameterized PM6 (rPM6) for hydrogen abstraction reactions. *Mol. Phys.* **118**, e1700313 (2020).

(27) Zhao, Y., Ng, H. T., Peverati, R. & Truhlar, D. G. Benchmark database for ylidic bond dissociation energies and its use for assessments of electronic structure methods. *J. Chem. Theory Comput.* **8**, 2824–2834 (2012).

(28) Wen, M., Blau, S. M., Spotte-Smith, E. W. C., Dwaraknath, S. & Persson, K. A. BonDNet: a graph neural network for the prediction of bond dissociation energies for charged molecules. *Chem. Sci.* **12**, 1858–1868 (2021).

(29) Qu, X., Latino, D. A. R. S. & Aires-De-sousa, J. A big data approach to the ultra-fast prediction of DFT-calculated bond energies. *J. Cheminform.* **5**, 34 (2013).

(30) St. John, P. C., Guan, Y., Kim, Y., Kim, S. & Paton, R. S. Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat. Commun.* **11**, 2328 (2020).

(31) Montgomery Jr., J. A., Frisch, M. J., Ochterski, J. W. & Petersson, G. A. A complete basis set model chemistry. VI. Use of density functional geometries and frequencies. *J. Chem. Phys.* **110**, 2822–2827 (1999).

(32) Montgomery Jr., J. A., Frisch, M. J., Ochterski, J. W. & Petersson, G. A. A complete basis set model chemistry. VII. Use of the minimum population localization method. *J. Chem. Phys.* **112**, 6532–6542 (2000).

(33) Wood, G. P. F. *et al.* A restricted-open-shell complete-basis-set model chemistry. *J. Chem. Phys.* **125**, 094106 (2006).

(34) Jarvis, M. W. *et al.* Direct detection of products from the pyrolysis of 2-phenethyl phenyl ether. *J. Phys. Chem. A* **115**, 428–438 (2011).

(35) Hudzik, J. M., Bozzelli, J. W. & Simmie, J. M. Thermochemistry of C_7H_{16} to $C_{10}H_{22}$ alkane isomers: Primary, secondary, and tertiary C-H bond dissociation energies and effects of branching. *J. Phys. Chem. A* **118**, 9364–9379 (2014).

(36) Menon, A. S., Wood, G. P. F., Moran, D. & Radom, L. Bond dissociation energies and radical stabilization energies: An assessment of contemporary theoretical procedures. *J. Phys. Chem. A* **111**, 13638–13644 (2007).

(37) Feng, Y., Liu, L., Wang, J.-T., Huang, H. & Guo, Q.-X. Assessment of experimental bond dissociation energies using composite ab initio methods and evaluation of the performances of density functional methods in the calculation of bond dissociation energies. *J. Chem. Inf. Comput. Sci.* **43**, 2005-2013 (2003).

(38) Otero-de-la-Roza, A. & DiLabio, G. A. Improved basis-set incompleteness potentials for accurate density-functional theory calculations in large systems. *J. Chem. Theory Comput.* **16**, 4176–4191 (2020).

(39) Prasad, V. K., Otero-de-la-Roza, A. & DiLabio, G. A. Atom-centered potentials with dispersion-corrected minimal-basis-set Hartree–Fock: An efficient and accurate computational approach for large molecular systems. *J. Chem. Theory Comput.* **14**, 726–738 (2018).

(40) Otero-de-la-Roza, A. & DiLabio, G. A. Transferable atom-centered potentials for the correction of basis set incompleteness errors in density-functional theory. *J. Chem. Theory Comput.* **13**, 3505–3524 (2017).

(41) *Comprehensive Handbook of Chemical Bond Dissociation Energies*, ed. Luo, Y.-R. (Taylor & Francis Group, LLC, 2007).

(42) Hanwell, M. D. *et al.* Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminform.* **4**, 17 (2012).

(43) https://www.ccdc.cam.ac.uk/solutions/csd-enterprise/applications/conformer-generator/.

(44) https://github.com/bobbypaton/FullMonte.

(45) Frisch, M. J. *et al.* Gaussian 16, Revision B.01 (Gaussian, Inc., Wallingford CT, 2016).

(46) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98**, 5648–5652 (1993).

(47) Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789 (1988).

(48) Vosko, S. H., Wilk, L., Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **58**, 1200–1211 (1980).

(49) Stephens, P. J., Devlin, F. J., Chabalowski, C. F. & Frisch, M. J. Ab Initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **98**, 11623–11627 (1994).

(50) Becke, A. D. A new mixing of Hartree-Fock and local density-functional theories. *J. Chem. Phys.* **98**, 1372–1377 (1993).

(51) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38**, 3098–3100 (1988).

(52) Hariharan, P. C. & Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor. Chim. Acta* **28**, 213–222 (1973).

(53) Francl, M. M. *et al.* Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *J. Chem. Phys.* **77**, 3654–3665 (1982).

(54) Goerigk, L. A comprehensive overview of the DFT-D3 London-dispersion correction. In *Non-Covalent Interactions in Quantum Chemistry and Physics: Theory and Applications*, eds. Otero-de-la-Roza & DiLabio G. A. (Elsevier Inc., 2017).

(55) Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132**, 154104 (2010).

(56) Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465 (2011).

(57) Johnson, E. R. & Becke, A. D. A post-Hartree-Fock model of intermolecular interactions: Inclusion of higher-order corrections. *J. Chem. Phys.* **124**, 174104 (2006).

(58) Yanai, T., Tew, D. P. & Handy, N. C. A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **393**, 51–57 (2004).

(59) Schäfer, A., Huber, C. & Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J. Chem. Phys.* **100**, 5829–5835 (1994).

(60) Johnson, E. R., Otero-de-la-Roza, A. & Dale, S. G. Extreme density-driven delocalization error for a model solvated-electron system. *J. Chem. Phys.* **139**, 184116 (2013).

(61) Roothaan, C. C. J. New developments in molecular orbital theory. *Rev. Mod. Phys.* **23**, 69–89 (1951).

(62) Pople, J. A. & Nesbet, R. K. Self-consistent orbitals for radicals. *J. Chem. Phys.* **22**, 571–572 (1954).

(63) *AB INITIO Molecular Orbital Theory*, eds. Hehre, W. J., Radom, L., Schleyer, P. v. R. & Pople, J. (Wiley, 1986)

(64) Lauderdale, W. J., Stanton, J. F., Gauss, J., Watts, J. D. & Bartlett, R. J. Many-body perturbation theory with a restricted open-shell Hartree-Fock reference. *Chem. Phys. Lett.* **187**, 21–28 (1991).

(65) Knowles, P. J., Andrews, J. S., Amos, R. D., Handy, N. C. & Pople, J. A. Restricted Møller-Plesset theory for open-shell molecules. *Chem. Phys. Lett.* **186**, 130–136 (1991).

(66) Lauderdale, W. J., Stanton, J. F., Gauss, J., Watts, J. D. & Bartlett, R. J. Restricted open-shell Hartree-Fock-based many-body perturbation theory: Theory and application of energy and gradient calculations. *J. Chem. Phys.* **97**, 6606–6620 (1992).

(67) Bartlett, R. J. Coupled-cluster theory: An overview of recent developments. In *Mordern Electronic Structure Theory Part II*, ed. Yarkony, D. R. (World Scientific, 1995).

(68) Prasad, V. K., Khalilian, H., Otero-de-la-Roza, A. & DiLabio, G. A. BSE49, a diverse, high-quality benchmark dataset of separation energies of chemical bonds. figshare https://doi.org/10.6084/m9.figshare.14544060.v1 (2021).

(69) Abdel-Rahman, M. A. *et al.* Mechanistic insights of the degradation of an O-anisidine carcinogenic pollutant initiated by OH radical attack: theoretical investigations. *New J. Chem.* **45**, 5907–5924 (2021).

(70) Nguyen, T. D.-T., Pham, N., Mai, T. V.-T., Nguyen, H. M. & Huynh, L. K. Detailed kinetic mechanism of thermal decomposition of furyl radicals: Theoretical insights. *Fuel* **288**, 119699 (2021).

(71) Chen, Y. F. *et al.* Insights into evolution mechanism of PAHs in coal thermal conversion: A combined experimental and DFT study. *Energy* **222**, 119970 (2021).

(72) Wang, Q.-D., Sun, M.-M. & Liang, J. Theoretical study of the hydrogen abstraction reactions from substituted phenolic species. *Comput. Theor. Chem.* **1196**, 113120 (2021).

(73) Liao, Z., Zeng, M. & Wang, L. Atmospheric oxidation mechansim of polychlorinated biphenyls (PCBs) initiated by OH radicals. *Chemosphere* **240**, 124756 (2020).

(74) Wang, L. & Wang, L. Atmospheric oxidation mechanism of acenaphthene initiated by OH radicals. *Atmos. Environ.* **243**, 117870 (2020).

(75) Sun, W., Hamadi, A., Abid, S., Chaumeix, N. & Comandini, A. An experimental and kinetic modeling study of phenylacetylene decomposition and the reactions with acetylene/ethylene under shock tube pyrolysis conditions. *Combust. Flame* **220**, 257–271 (2020).

(76) Zeng, M., Liao, Z. & Wang, L. Atmospheric oxidation of gaseous anthracene and phenanthrene initiated by OH radicals. *Atmos. Environ.* **234**, 117587 (2020).

(77) Wang, S. *et al.* Aromatic photo-oxidation, a new source of atmospheric acidity. *Environ. Sci. Technol.* **54**, 7798–7806 (2020).

(78) Abdel-Rahman, M. A., Shibl, M. F., El-Demerdash, S. H. & El-Nahas, A. M. Simulated kinetics of the atmospheric removal of aniline during daytime. *Chemosphere* **255**, 127031 (2020).

(79) Khojandi, M., Seif, A., Zahedi, E., Domingo, L. R. & Karimkhani, M. Unravelling the kinetics and molecular mechanism of the degenerate Cope rearrangement of bullvalene. *New J. Chem.* **44**, 6543–6552 (2020).

(80) Padash, R. Mechanism and kinetic investigations of 5-fluorouracil tautomeric conversions in the gas phase: DFT and CBS-QB3 methods using multichannel Rice–Ramsperger–Kassel–Marcus steady-state approximation theory. *Theor. Chem. Acc.* **139**, 1–9 (2020).

(81) Doroshenko, I., Vaskivskyi, Y. & Chernolevska, Y. Structural transformations in solid and liquid n-butanol from FTIR spectroscopy. *Mol. Cryst. Liq. Cryst.* **697**, 11–19 (2020).

(82) Ning, H., Wu, J., Ma, L. & Ren, W. Exploring the pyrolysis chemistry of prototype aromatic ester phenyl formate: Reaction pathways, thermodynamics and kinetics. *Combust. Flame* **211**, 337–346 (2020).

(83) Poskrebyshev, G. A. The CBS values of $\Delta_fH^o_{298.15}$ and $S^o_{298.15}$ of the phenoxy radicals, formed by abstraction of H atom from the components of surrogate bio-oil. *Comput. Theor. Chem.* **1169**, 112625 (2019).

(84) Abdel-Rahman, M. A., Shibl, M. F., El-Demerdash, S. H. & El-Nahas, A. M. First-principle studies on the gas phase OH-initiated oxidation of O-toluidine. *Comput. Theor. Chem.* **1170**, 112634 (2019).

(85) Wagner, J. P. Gauging stability and reactivity of carbonyl: O-oxide Criegee intermediates. *Phys. Chem. Chem. Phys.* **21**, 21530–21540 (2019).

(86) Bietti, M. *et al.* Evaluation of polar effects in hydrogen atom transfer reactions from activated phenols. *J. Org. Chem.* **84**, 1778–1786 (2019).

(87) Vaskivskyi, Y. *et al.* 1-Hexanol conformers in a nitrogen matrix: FTIR study and high-level ab initio calculations. *J. Mol. Liq.* **278**, 356–362 (2019).

(88) Bain, M., Hansen, C. S., Karsili, T. N. V. & Ashfold, M. N. R. Quantifying rival bond fission probabilities following photoexcitation: C-S bond fission in t-butylmethylsulfide. *Chem. Sci.* **10**, 5290–5298 (2019).

(89) Poskrebyshev, G. A. The values of $\Delta_fH^o_{298.18}(Al_nO_mH_p)$, calculated using the CBS correction dependencies, as well as the thermochemistry of the isodesmic/homodesmotic reactions. *Comput. Theor. Chem.* **1164**, 112540 (2019).

(90) DiLabio, G. A. *et al.* Hydrogen atom transfer (HAT) processes promoted by the Quinolinimide-*N*-oxyl radical. A kinetic and theoretical study. *J. Org. Chem.* **82**, 6133–6141 (2017).

Chapter 6

BH9, a new comprehensive benchmark dataset for barrier heights and reaction energies: Assessment of density functional approximations and basis set incompleteness potentials

Abstract

The calculation of accurate reaction energies and barrier heights is essential in computational studies of reaction mechanisms and thermochemistry. In order to assess methods regarding their ability to predict these two properties, high-quality benchmark sets are required that comprise a reasonably large and diverse set of organic reactions. Due to the time-consuming nature of both locating transition states and computing accurate reference energies for reactions involving large molecules, previous benchmark sets have been limited in scope, the number of reactions considered, and the size of the reactant and product molecules. Recent advances in coupled-cluster theory, in particular local correlation methods like DLPNO-CCSD(T), now allow the calculation of reaction energies and barrier heights for relatively large systems. In this work, we present a comprehensive, and diverse benchmark set of barrier heights and reaction energies based on DLPNO-CCSD(T)/CBS, called BH9. BH9 comprises 449 chemical reactions belonging to nine types common in organic chemistry and biochemistry. We examine the accuracy of DLPNO-CCSD(T) vis-a-vis canonical CCSD(T) for a subset of BH9 and conclude that, although there is a penalty in using the DLPNO approximation, the reference data are accurate enough to serve as benchmark for density-functional theory (DFT) methods. We then present two applications of the BH9 set. First, we examine the performance of several density functional approximations commonly used in thermochemical and mechanistic studies. Second, we assess our basis set incompleteness potentials regarding their ability to mitigate basis set incompleteness error. The number of data points, the diversity of the reactions considered, and the relatively large size of the reactant molecules make BH9 the most comprehensive thermochemical benchmark set to date, and a useful tool for the development and assessment of computational methods.

1. Introduction

The prediction of barrier heights (BHs) and reaction energies (REs) using computational methods, combined with the application of transition-state theory,[1–3] is a powerful tool for the elucidation of reaction mechanisms in chemistry.[4,5] The prediction of kinetic and thermochemical properties is also important in biochemistry, and has contributed greatly to the understanding of the catalytic activity of enzymes,[6–9] as well as to the discovery of new drugs.[10,11]

The main bottleneck for the successful prediction of rate constants and equilibrium constants is the accuracy in the determination of BHs and REs.[3,10] Because of the exponential dependence of these constants on the corresponding energies, an accuracy of about RT (0.6 kcal/mol at room temperature) or better is required.[12] Quantum mechanical methods based on wavefunction theory,[12–15] particularly recent composite methods, are able to calculate BHs and REs to this level of accuracy,[16–18] but they are not applicable to molecules with sizes typically encountered in organic chemistry, let alone biochemistry.[3] As a consequence of the trade-off between accuracy and computational cost, the most popular method for thermochemical and kinetic calculations in organic reactions is density-functional theory (DFT).[10] In reactions of biochemical interest, where the reactant molecules are much larger, DFT is typically combined with force fields in hybrid quantum mechanics/molecular mechanics (QM/MM) approaches.[6,7,19] In either case, the accuracy of the methods typically employed is often sufficient for gauging the relative energies of various mechanistic pathways but not enough to reliably predict rate constants of chemical reactions.[10,20] Consequently, the search for a standard method for kinetic and mechanistic studies is still ongoing.[4,5]

To develop new computational methods for the study of chemical reactions, and to assess the existing ones, high-quality benchmark sets are necessary.[21–27] These benchmark sets comprise REs and BHs of model reactions calculated at a very accurate level of theory, typically coupled-cluster theory (CC) with large basis sets and a complete-basis-set (CBS) extrapolation[16] (CCSD(T)/CBS is a very popular method[12]). Besides the obvious requirement that the reference data be accurate, there are a number of additional desirable traits for BH and RE benchmark sets. First, the set of reactions must be sufficiently large for the analysis to be statistically significant, and diverse enough to catch any particularities or biases of the method under study. For instance, most DFT methods tend to underestimate BH of pericyclic reactions because the transition state (TS) is over stabilized due to delocalization error.[28,29] Second, non-covalent interactions between reactants play an important role in stabilizing the TS.[30] The importance of this stabilization increases with the size of the reactant molecules and it is particularly important in biochemical studies where, for instance, the shape of the active site determines the activity and specificity of enzymes.[4] Therefore, it is essential that the reactant molecules in the benchmark set are large enough to correctly assess the method under study regarding its ability to describe non-covalent interactions.[19,30–33]

There are difficulties with the creation of benchmark sets for BHs and REs with the aforementioned characteristics. The generation of TSs is not easily automatized.[10,20] More importantly, the computational cost involved in the calculation of accurate reference data limits the number of reactions in the set and the size of the reactant molecules. As a consequence, previously proposed benchmark sets use model reactions with small reactant molecules that are not representative of the typical reactions commonly found in mechanistic studies.[34–37] Other benchmark sets either focus on specific types of reactions, or they contain only a handful of data points, or they are not evaluated using a reference level of enough quality to allow benchmarking commonly used quantum mechanical methods.[29,32,33,38–44] The current necessity of a benchmark set for enzymatically catalyzed reactions has been emphasized several times recently.[32,33,45]

Local correlation methods, particularly DLPNO-CCSD(T), have become very popular recently due to a favorable combination of relatively high accuracy and modest computational cost.[46–51] Thanks to its near-linear-scaling nature, DLPNO-CCSD(T) can be applied to reasonably large systems.[19] Since conventional CCSD(T)/CBS is at least two orders of magnitude more accurate than the methods typically assessed with BH and RE benchmark sets, a trade-off is used in this work. By using DLPNO-CCSD(T)/CBS for the reference energies, we designed a benchmark set (called BH9) that has the desirable features listed above, namely, the reactions in BH9 are numerous and diverse and the reactants are relatively large. The accuracy penalty in using the DLPNO approximation[31,52] is evaluated, providing an accuracy limit for the assessment of approximate methods. Variants of the DLPNO-CCSD(T)/CBS approach have been used in other recently proposed benchmark sets.[32,33,53,54]

To our knowledge, BH9 is the most comprehensive benchmark set for BHs and REs of organic and bio-organic reactions to date. Our particular objective with this set is to aid in the development of atom-centered potentials[55–57] (ACPs), whose training requires a large and diverse set of molecular properties. However, recent machine-learning-based methods can equally benefit from using the BH9 data. Furthermore, the reference BHs and REs in BH9 can be recalculated should further developments in computational methods or computer hardware occur, without the need to find TSs for new reactions, a task that is often non-trivial.

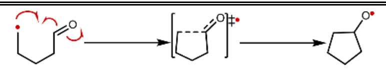
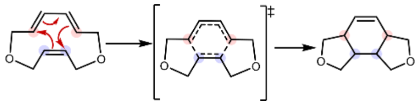
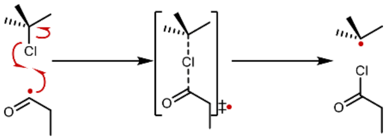
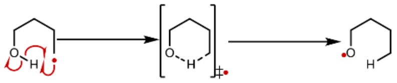
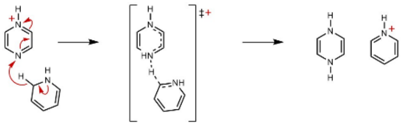
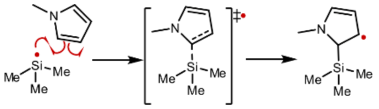
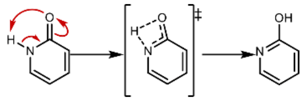
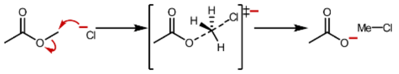
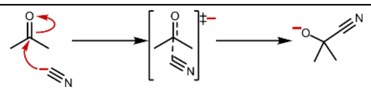
We also present two simple applications of the new benchmark set. First, we use BH9 to assess several popular density functional approximations used in mechanistic studies. The effect of including corrections for dispersion interactions is considered, and we analyze the performance of these functionals individually for the different types of reactions included in the BH9 set. Second, the application of DFT to reaction mechanisms in practice often requires using a finite basis set due to computational constraints. Therefore, we also study the performance of our basis set incompleteness potentials[55,56] (BSIPs) regarding their ability to mitigate basis set incompleteness error in the calculation of REs and BHs.

2. Design of BH9 and Computational Details

2.1 Design of the BH9 benchmark set

The BH9 set contains 449 elementary chemical reactions, categorized in the reaction types shown in Table I. The reference data comprises the corresponding 449 REs and 898 BHs (forward and reverse), as well as the structures of reactants, products, and transition states. Table I also shows a prototype reaction for each type. The full list of diagrams for each reaction is given in the Supporting Information (SI), as well as the reference BHs, REs, and the geometries of all the molecular species. The data for each reaction is given in the form of "db" files. This plain-text file format has been described elsewhere.[58,59] A representative subset of the BH9 set can be statistically derived using, for instance, the technique described in Ref. 60.

Table I. Reaction types in the BH9 set.

Reaction type		Number[a]	Example reaction[b]
I	Radical rearrangement and addition	48	
II	Pericyclic	140	
III	Halogen atom transfer	43	
IV	Hydrogen atom transfer	90	
V	Hydride transfer	42	
VI	B- and Si- containing reactions	35	
VII	Proton transfer	10	
VIII	Nucleophilic substitution	15	
IX	Nucleophilic addition	26	

[a] number of reactions in each type, [b] example reaction for each type

The reaction types in Table I represent a diverse set of reactions that are common in organic and bio-organic chemistry, although the list is by no means exhaustive. Most reaction types and many of the particular reactions included in the BH9 set were adopted from the Mechanism and Catalytic Site Atlas (M-CSA) database,[61] and, are known to occur in biological systems. However, some reaction types that are important in organic chemistry, such as pericyclic, hydride-transfer, and halogen-atom transfer reactions, are relatively rare in biological contexts or are not sufficiently represented in the M-CSA. For these reaction types, we explored the literature and compiled a number of reactions from various published mechanistic studies in order to complete our database.[62–106]

For the sake of simplicity, and due to our desire for this set to serve as a basis for ACP development, all molecules in BH9 contain exclusively elements common in organic chemistry (H, C, N, O, F, P, S, Cl). We also included a specific set of reactions containing Si and B (reaction type VI in Table I). The fact that there are no transition metals in the BH9 reactions simplifies the application and interpretation of the tests based on this set, particularly regarding the application of BH9 to the assessment of DFT methods. RE and BH benchmark sets for reactions containing transition metals have been proposed recently,[53,54,107,108] some of them also at DLPNO-CCSD(T) level.[53,54] The sizes of the reactant and TS molecules in BH9 range from 11 to 71 atoms—significantly larger than most previous sets, and typical of mechanistic studies.

Some of the reactions in BH9, particularly nucleophilic substitutions, nucleophilic additions, and proton transfer reactions, involve charged species. In this case, we expect the species involved in the reaction, and particularly reactants and products, to be greatly stabilized by interactions with the solvent or the environment. We experienced difficulties finding some of these TSs, which is why the number of reactions in these three categories is smaller than the others (see Table I). In addition, some of the BHs are negative, possibly because the solvent stabilizes reactants and products more than it stabilizes the TS. Although we eliminated very negative BHs from the set, some were left for diversity sake. A similar decision was taken by Iron *et al.* for their BH set for reactions involving transition metals.[53]

2.2 Location of the transition states

Guess TSs were built for the 449 reactions in the BH9 set. This was a laborious process because of the difficulty in locating TS with the currently available algorithms in standard software packages. In addition to not being automatic, the TS search often failed entirely, which explains the uneven number of reactions in each category of Table I. Because of their relatively large size and the abundance of reactions, reliably locating the minimum-energy conformer for each species in the BH9 is a formidable problem.

However, in order for the BHs to still be representative of the corresponding reactions, we devised a protocol that explores the conformational landscape of reactants, products, and TS. This protocol, described below, ensures that the proposed structures are reasonably close in energy, if not identical, to the global energy minima of all species.

In all cases, we used the Gaussian 09/16[109,110] software package. Our calculations employed a default SCF convergence criterion of 10^{-8} Hartree, "ultrafine" integration grid (pruned 99 × 590 grid), and tight optimization convergence criteria (maximum force = 1.5×10^{-5} Hartree/Bohr, RMS force = 1×10^{-5} Hartree/Bohr, maximum displacement = 6×10^{-5} Bohr, RMS displacement = 4×10^{-5} Bohr). The CalcFC and NoEigenTest options were used to specify the computation of force constants in the first step of the optimization and to suppress the curvature test during optimization, respectively. All calculations were carried out in the gas phase.

In the first step, preliminary TS were located by geometry optimization followed by a frequency calculation. Finding the TS is often difficult because a good initial guess for the TS geometry is required for the optimization to succeed. In difficult cases, we ran series of constrained geometry minimizations where we fixed a few geometric parameters, then used the resulting structure as the initial guess for the TS search. The preliminary TS optimizations used the B3LYP hybrid density functional,[111,112] except in a few cases where the range-separated density functional CAM-B3LYP[113] was used. (The change in functional was prompted by the instability of B3LYP in the calculation of zwitterionic systems.) The D3 dispersion correction[114] with Becke–Johnson damping[115,116] was used in all cases. Due to the different sizes of the reactant molecules, depending on the reaction type, various Pople basis sets[117–119] (6-31G*, 6-31+G*, 6-31+G**) were used together with their associated basis set incompleteness potentials[56] (BSIPs) to mitigate the effect of basis set incompleteness error (BSIE). 6-31G*-BSIP was used to model radical addition and pericyclic reactions. 6-31+G*-BSIP was used to model halogen atom transfer, nucleophilic substitution, nucleophilic addition, and the B- and Si-containing reactions. 6-31+G**-BSIP was used to model hydrogen atom transfer, hydride transfer, and proton transfer reactions. Each preliminary TS was checked for the presence of a single imaginary frequency and visually inspected to confirm the imaginary-frequency eigenvector was oriented along the reaction coordinate.

The preliminary TS were then subjected to a constrained conformer search using the commercial Schrödinger's MacroModel Suite[120,121] implemented in the Maestro[122] software package. This search is similar to the one used in previous works.[123–125] Bonds undergoing breaking and formation in the TS had their bond distances constrained to their values in the preliminary TS. The conformational search was performed using the mixed torsional/large-scale low-mode sampling option in Maestro, followed by a constrained post-optimization with the OPLS all-atom force field.[126] From this sampling, a maximum of 100 structures (fewer if the molecule was not sufficiently flexible) were then subjected to a single-point calculation using the same calculation level as in the preliminary TS optimization. A maximum of 9 lowest-energy conformers were chosen to undergo further refinement.

For each reaction, the nine TS conformers obtained in this manner plus the TS from the preliminary optimization were subjected to unconstrained optimization using the same method as above. We discarded all the structures whose optimization failed to locate a new TS or whose eigenvectors did not point in the direction of reactants and products. The lowest-energy conformer was then subjected to a final TS optimization and frequency calculation at a higher level of theory (CAM-B3LYP-D3(BJ)/6-311++G**[127,128]). After verification of the imaginary frequencies and the direction of the imaginary-frequency eigenvector, this last structure was adopted as the TS for the reaction.

2.3 Reactant and product structures

The initial reactant and product structures were constructed from the optimized TS and subjected to geometry optimizations using CAM-B3LYP-D3(BJ) with the same combination of Pople basis sets and BSIPs as above (6-31G*, 6-31+G*, or 6-31+G** depending on reaction type). All geometry optimizations employed a default SCF convergence criterion of 10^{-8} Hartree, "ultrafine" integration grid, and the default optimization convergence criteria (maximum force = 4.5×10^{-4} Hartrees/Bohr, RMS force = 3×10^{-4} Hartrees/Bohr, maximum displacement = 1.8×10^{-3} Bohr, RMS displacement = 1.2×10^{-3} Bohr).
After this initial relaxation, a 100-step Monte-Carlo multiple minimum[129] (MCMM) conformational search was carried out using the FullMonte[130,131] software package. The conformers generated in this way were optimized with the semi-empirical PM6-DH2[132] method using the MOPAC2016[133] software package. All conformers were then subjected to a single point calculation at the same level of theory used for their initial optimization. The ten lowest-energy conformers were selected for further optimization at the same level. The resulting lowest energy conformer was subjected to a final optimization and frequency calculation at a higher level of theory (CAM-B3LYP-D3(BJ)/6-311++G**).

2.4 Reference energy calculations

The reference BHs and REs were obtained using single-point DLPNO-CCSD(T)[46–51] (in particular, DLPNO-CCSD(T0)) at the equilibrium geometries of reactants, products, and TSs calculated as above. The favorable scaling of DLPNO-CCSD(T) makes it possible to apply CC to the fairly large molecules included in BH9, which is why this method has been often used to generate reference data in recent benchmark sets.[32,53,54] Naturally, the use of the DLPNO approximation introduces an error compared to canonical CCSD(T). The reference energies are calculated using a focal-point approach to minimize the computational cost associated with using large basis sets.[12] The error introduced by DLPNO as well as the convergence of the reference data with respect to basis set size are examined in the Results and Discussion section.

For the calculation of the reference energies, we used the ORCA program, version 4.2.1.[134,135] The aug-cc-pVNZ basis sets (in the following, aNZ for short) of Dunning and co-workers[136–138] were used for the complete-basis-set extrapolation, as well as the resolution of the identity MP2 method[139–141] (RI-MP2) with the aug-cc-pVNZ/C auxiliary basis sets.[142] The TightPNO and TightSCF threshold settings were used in the DLPNO-CCSD(T) calculation. The use of TightPNO was shown to be very important in the calculation of REs and BHs, particularly those of Diels-Alder reactions.[31] The frozen core approximation was used in all calculations. It has been shown to have a relatively minor impact on the accuracy of calculated thermochemical properties.[12]

2.5 DFT calculations details

DFT calculations were used to assess the performance of various density functional approximations commonly used in mechanistic studies[10] on the BH9 set. We used Gaussian 16[110] to calculate the BH9 reactions using B3LYP,[111,112] LC-ωPBE,[143,144] M05-2X,[145] M06-2X,[146] revTPSS,[147] and ωB97XD.[148,149] Ultrafine grids were used for all calculations. The BLYP,[112,150] PBE,[151] TPSS,[152] BH&HLYP,[112,153] PBE0,[154] and CAM-B3LYP[113] functionals were evaluated using ORCA, version 4.2.1.[134,135] The tight SCF convergence criteria and the “grid4” integration grid were used. Second-order SCF was deactivated. The resolution of the identity (RI) method was used in all cases. For the hybrid functionals (BH&HLYP and PBE0), RI was applied to both the Coulomb and exchange integrals (RI-JK keyword). For the range-separated hybrid functional (CAM-B3LYP), the chain-of-spheres approximation[155] was used to calculate the exchange energy. In all cases, the Def2-QZVPP basis set was used,[156] with the corresponding auxiliary basis sets (Def2/JK and Def2/J) used where appropriate.[157] Contrary to previous reports,[158] we did not observe any SCF convergence problems using the Minnesota functionals.

To evaluate the importance of dispersion, the exchange-hole dipole moment (XDM) model was used in combination with some of the functionals above.[159,160] The canonical complete-basis-set XDM damping function parameters and the postg program were used.[161,162] We expect the conclusions from this analysis to be transferable to other dispersion corrections, such as Grimme's Dn family.[114]

Due to the typical size of the molecular species involved, the availability of computationally inexpensive methods for thermochemistry and kinetics is very important in the study of biochemical reactions.[57] One of the major factors impacting the accuracy of DFT methods in this context is BSIE, which arises from the finite nature of the basis sets employed. For this reason, we also examine the performance of our recently proposed basis set incompleteness potentials[56] (BSIPs) combined with several double-ζ basis sets in the description of REs and BHs. In particular, we evaluated the PBE0-XDM functional[154,160] combined with the BSIP-corrected 6-31G*, 6-31+G*, 6-31+G**,[117–119] Def2-SV(P), Def2-SVP,[156] and pc-1 basis sets.[163–166]

3. Results and Discussion

3.1 Evaluation of the reference data

The most popular calculation level for benchmark sets is CCSD(T) with CBS extrapolation, which is known to yield sub-kcal/mol accuracy.[12,167] As mentioned above, this level of theory is too computationally demanding and cannot be used to generate reference data for BH9, so we opted for DLPNO-CCSD(T) instead. We expect the two primary sources of error are our choice of basis set extrapolation strategy[12] and the application of the DLPNO approximation.[31] In this section, we evaluate the importance of both sources of error, and we provide a reasonable estimate for the error bars associated with the BH9 reference data. Ultimately, this error estimate constitutes the accuracy limit of the BH9 set; methods more accurate than those applied here cannot be reliably assessed with this set.

The reference data was calculated using a focal-point approach,[168,169] which has been shown to be an effective way of approaching the CBS limit in similar calculations.[12,170,171] The BH9 reference energies are calculated as:

$$E = E_{HF}^{a\{T,Q\}Z} + \Delta E_{MP2}^{a\{T,Q\}Z} + \Delta\, E_{DLPNO-CCSD(T)}^{aTZ} \tag{1}$$

where $E_{HF}^{a\{T,Q\}Z}$ is the HF energy calculated from the aTZ and aQZ energies using the CBS extrapolation formula:

$$E_{HF}^{L} = E_{HF}^{CBS} + Aexp(-\alpha\sqrt{L}) \tag{2}$$

where L is the cardinal number of the basis set (3 for aTZ, 4 for aQZ, etc.). From this formula, a two-point extrapolation approach can be easily derived:

$$E_{HF}^{CBS} = \frac{E_{HF}^{X} \times e^{-\alpha\sqrt{Y}} - E_{HF}^{Y} \times e^{-\alpha\sqrt{X}}}{e^{-\alpha\sqrt{Y}} - e^{-\alpha\sqrt{X}}} \tag{3}$$

where X and Y are the cardinal numbers of the basis set pair. Following the recommendations of Neese and Valeev, we used the optimized α = 5.79 value for the aTZ/aQZ pair.[172]

The MP2 correlation energy (ΔE_{MP2}) is calculated using the known inverse cube dependence of the correlation energy with the basis set cardinal number:[173,174]

$$\Delta E_{MP2}^{L} = \Delta E_{MP2}^{CBS} + AL^{-\beta} \tag{4}$$

with β = 3 in the large-L limit. This yields the two-point extrapolation formula:

$$\Delta E_{MP2}^{CBS} = \frac{E_{MP2}^{X} \times X^{\beta} - E_{MP2}^{Y} \times Y^{\beta}}{X^{\beta} - Y^{\beta}} \tag{5}$$

In practice, optimized β parameters have been proposed for some basis set pairs, and it has been shown that β < 3 improves the CBS estimate for low cardinal numbers.[175] The β = 3.05 value proposed by Neese and Valeev for the aTZ/aQZ pair is used here.[172] Finally, the last component in our reference energy is:

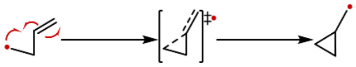
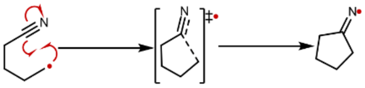
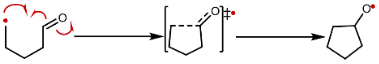
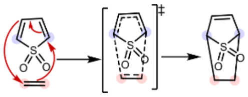
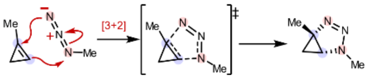
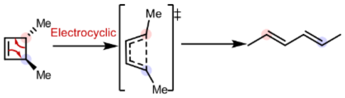
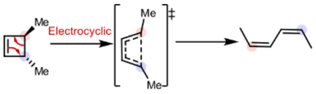
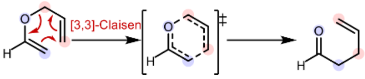
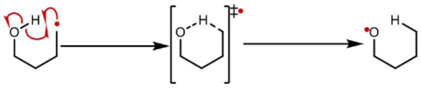
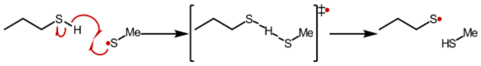
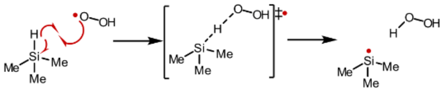
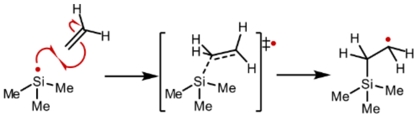
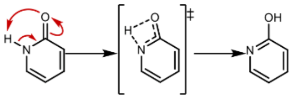
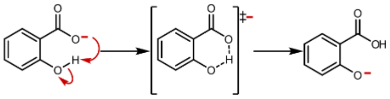
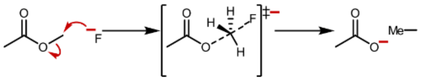
$$\Delta E_{DLPNO-CCSD(T)}^{aTZ} = E_{DLPNO-CCSD(T)}^{aTZ} - E_{MP2}^{aTZ} \tag{6}$$

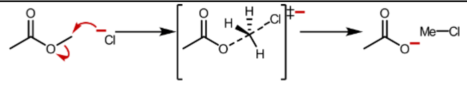
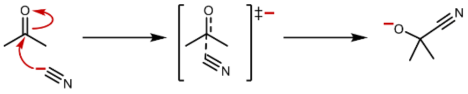
which is calculated using the aTZ basis set. The CCSD(T)/MP2 energy difference is routinely calculated at the aTZ level in the “gold standard” focal-point approach for non-covalent interactions,[176–178] and it is justified by the observation that high-order contributions to the correlation energy converge relatively quickly with the basis set size.[167] In fact, our method for the calculation of reference data is very similar to the “gold standard” method for intermolecular interactions, except for the use of DLPNO for the CCSD(T) calculation.

To estimate the overall error in the BH9 reference data and to assess each of the approximations made, we selected a small subset of BH9 containing 17 reactions with relatively small molecules. This subset is shown in Table II. Since errors in the calculation of REs are typically lower than BHs,[31] we focus on the latter. The small size of the molecules in this subset allows the calculation of canonical

CCSD(T)/aTZ energies, as well as HF/a5Z and MP2/a5Z. These last two quantities permit the calculation of the extrapolated aQZ/a5Z HF and MP2 barrier heights. For the MP2 correlation energy, we used the same two-point extrapolation formula (Eq. 5) with the asymptotic value $\beta = 3$.[173,174] For the HF energy, we used the extrapolation formula proposed by Karton and Martin for this particular basis set pair.[179,180]

Table II. Subset of the BH9 reactions used for assessing the quality of the reference data.

#	Reaction	Type
1		Radical rearrangement (I)
2		Radical rearrangement (I)
3		Radical rearrangement (I)
4		Pericyclic (II)
5		Pericyclic (II)
6		Pericyclic (II)
7		Pericyclic (II)
8		Pericyclic (II)
9		Hydrogen atom transfer (IV)
10		Hydrogen atom transfer (IV)
11		Si-containing (VI)
12		Si-containing (VI)
13		Proton transfer (VII)
14		Proton transfer (VII)
15		Nucleophilic substitution (VIII)

#	Reaction	Type
16		Nucleophilic substitution (VIII)
17		Nucleophilic addition (IX)

The BHs obtained with these methods are shown in Table III. We first consider the impact of basis set incompleteness on the individual components of our reference BHs. In the case of HF, our best CBS estimate (a{Q,5}Z extrapolation) agrees with HF/a5Z to within 0.01 kcal/mol on average, indicating that both are converged to within this value. Our chosen reference method for the HF component (a{T,Q}Z extrapolation) has a mean absolute error (MAE) of only 0.05 kcal/mol with respect to the aQZ/a5Z result. The highest deviations happen for reactions involving second-row atoms: numbers 6 (0.10 kcal/mol, both directions), 4 (0.15 kcal/mol, both directions), 8 (0.19 kcal/mol, reverse), and 12 (0.25 kcal/mol, reverse).

Table III. Barrier heights for the reactions in Table II calculated using various levels of theory.[a]

		HF					ΔMP2[f]					ΔCCSD(T)[f]		CCSD(T)	
Reaction[g]		aTZ[b]	aQZ[b]	a5Z[b]	a{T,Q}Z[c]	a{Q,5}Z[c]	aTZ[b]	aQZ[b]	a5Z[b]	a{T,Q}Z[c]	a{Q,5}Z[c]	DLPNO aTZ[b]	Can.[d] aTZ[b]	DLPNO CBS[e]	Can.[d] CBS[e]
1	F	16.84	16.85	16.85	16.85	16.85	-1.25	-1.08	-1.27	-0.96	-1.47	-4.15	-3.76	11.74	12.13
	R	10.09	10.15	10.15	10.17	10.14	5.30	5.46	5.58	5.57	5.71	-6.74	-6.35	9.01	9.39
2	F	16.38	16.42	16.44	16.43	16.44	4.80	5.13	5.24	5.36	5.36	-11.00	-10.80	10.80	11.00
	R	26.39	26.42	26.41	26.43	26.41	-4.18	-4.03	-3.96	-3.92	-3.90	-1.98	-1.69	20.54	20.82
3	F	15.67	15.84	15.86	15.89	15.86	-0.96	-0.68	-0.55	-0.48	-0.42	-6.85	-6.78	8.56	8.63
	R	19.79	19.70	19.69	19.67	19.69	-6.12	-6.01	-5.98	-5.93	-5.94	-2.90	-2.62	10.84	11.13
4	F	42.42	43.16	43.46	43.36	43.51	-39.38	-38.45	-38.06	-37.78	-37.65	11.41	10.46	16.98	16.03
	R	70.71	70.48	70.29	70.41	70.26	-15.48	-15.46	-15.39	-15.44	-15.32	4.49	3.74	59.46	58.71
5	F	41.06	41.34	41.45	41.41	41.47	-35.95	-35.32	-35.19	-34.87	-35.06	8.89	8.02	15.43	14.56
	R	87.78	87.91	87.88	87.95	87.88	-32.15	-31.60	-31.34	-31.21	-31.06	11.55	10.69	68.29	67.44
6	F	41.85	41.97	41.97	42.00	41.97	-11.06	-10.93	-10.82	-10.83	-10.71	1.91	1.62	33.08	32.78
	R	59.96	60.12	60.15	60.16	60.15	-18.88	-18.66	-18.61	-18.50	-18.55	3.76	3.58	45.42	45.24
7	F	56.32	56.42	56.43	56.45	56.43	-14.28	-13.95	-13.81	-13.72	-13.66	1.55	1.12	44.28	43.84
	R	70.99	71.10	71.15	71.13	71.15	-20.69	-20.34	-20.25	-20.08	-20.16	3.30	3.01	54.35	54.06
8	F	50.55	50.77	50.83	50.83	50.84	-25.91	-25.53	-25.41	-25.27	-25.29	6.51	5.68	32.08	31.25
	R	69.02	69.36	69.39	69.44	69.40	-25.53	-25.43	-25.30	-25.36	-25.16	5.53	4.62	49.61	48.70
9	F	28.85	29.05	29.08	29.11	29.09	-9.04	-8.73	-8.60	-8.51	-8.48	-4.18	-4.47	16.42	16.13
	R	33.81	33.93	33.96	33.96	33.96	-22.61	-22.81	-22.80	-22.95	-22.80	1.36	1.06	12.36	12.06
10	F	21.25	21.42	21.55	21.47	21.57	-16.87	-16.83	-16.82	-16.81	-16.81	0.13	-0.36	4.80	4.30
	R	21.34	21.51	21.63	21.55	21.65	-17.17	-17.10	-17.08	-17.05	-17.06	0.08	-0.25	4.58	4.25
11	F	34.65	35.21	35.38	35.36	35.41	-22.00	-22.02	-21.91	-22.04	-21.79	2.40	1.44	15.72	14.76
	R	19.87	19.85	19.78	19.84	19.77	-4.60	-4.32	-4.16	-4.12	-3.98	-5.04	-5.70	10.68	10.02
12	F	8.54	8.64	8.66	8.66	8.66	-2.00	-1.60	-1.48	-1.32	-1.36	-5.11	-5.10	2.24	2.25
	R	30.06	30.47	30.72	30.57	30.76	3.11	3.41	3.57	3.62	3.74	-5.84	-5.80	28.35	28.39
13	F	49.98	50.22	50.24	50.28	50.25	-15.90	-16.08	-16.11	-16.20	-16.14	3.21	2.93	37.29	37.01
	R	51.08	51.28	51.31	51.33	51.32	-14.50	-14.64	-14.69	-14.75	-14.75	1.62	1.37	38.20	37.96
14	F	2.70	2.76	2.77	2.78	2.77	-2.24	-2.19	-2.16	-2.16	-2.12	0.22	0.26	0.84	0.87
	R	2.96	3.03	3.05	3.05	3.05	-2.95	-2.99	-2.96	-3.02	-2.93	0.72	0.75	0.75	0.78

		HF					ΔMP2^f					ΔCCSD(T)^f		CCSD(T)	
Reaction^g		aTZ^b	aQZ^b	a5Z^b	a{T,Q}Z^c	a{Q,5}Z^c	aTZ^b	aQZ^b	a5Z^b	a{T,Q}Z^c	a{Q,5}Z^c	DLPNO aTZ^b	Can.^{d} aTZ^b	DLPNO CBS^e	Can.^{d} CBS^e
15	F	3.80	3.93	3.99	3.97	4.00	-6.03	-5.94	-5.77	-5.88	-5.60	-1.29	-1.88	-3.21	-3.80
	R	25.57	25.87	26.01	25.95	26.03	-11.34	-11.27	-11.23	-11.23	-11.18	0.34	-0.34	15.07	14.39
16	F	25.57	25.71	25.73	25.75	25.73	-11.53	-11.24	-11.20	-11.04	-11.15	-0.42	-1.09	14.28	13.62
	R	2.72	3.28	3.63	3.43	3.68	-1.87	-1.90	-1.92	-1.91	-1.95	-1.56	-2.12	-0.05	-0.61
17	F	13.08	13.40	13.43	13.48	13.44	-16.69	-16.60	-16.42	-16.54	-16.22	2.11	1.68	-0.94	-1.38
	R	3.48	3.39	3.39	3.36	3.39	2.93	3.02	3.02	3.09	3.01	-1.64	-1.68	4.81	4.77
MAE^h		0.27	0.08	0.01	0.05		0.42	0.21	0.11	0.11		0.43		0.43	

a) units are kcal/mol, b) aNZ = aug-cc-pVNZ basis set, c) a{X,Y}Z = two-point HF or correlation energy extrapolation as described in the text, d) Can. = canonical, e) CBS = complete-basis-set extrapolation (Eq. 1), f) ΔMP2 = $E_{MP2} – E_{HF}$, ΔCCSD(T) = $E_{CCSD(T)} – E_{MP2}$, g) F = forward reaction, R = reverse reaction, h) MAEs are relative to the large basis set results of each method

As noted above, the MP2 correlation energy converges more slowly to the CBS than the HF energy so, as expected, the basis set incompleteness errors are higher. The MP2 correlation contribution used in our reference method (a{T,Q}Z extrapolation) has the same MAE as MP2/a5Z (0.11 kcal/mol) compared to our best MP2/CBS estimate (a{Q,5}Z extrapolation). In this case, the large errors are not associated with second-row atoms, and they can be as high as half a kcal/mol (0.51 kcal/mol for forward reaction 1 and 0.32 kcal/mol for forward reaction 13). Combining the HF and MP2 results, we expect the average error from the HF+MP2 contribution to be in the vicinity of 0.2 kcal/mol, with worst cases being between 0.5 and 1 kcal/mol. Due to computational constraints, we cannot estimate the error introduced by calculating $\Delta E_{CCSD(T)}$ at aTZ level, although past experience with non-covalent interactions suggests that it is in the range of tenths of a kcal/mol or lower.[176]

The last four columns in Table III show the error introduced by the DLPNO approximation by comparing the $\Delta E_{CCSD(T)}$ contribution and the total BH with and without DLPNO. The MAE from the DLPNO approximation is 0.43 kcal/mol, which is very similar to the 0.51 kcal/mol reported by Paiva *et al.* for enzymatic reactions.[52] However, there are a few reactions where the deviations between DLPNO and canonical CCSD(T) are significantly higher, although lower than 1 kcal/mol in all cases: reactions 17 (0.83 kcal/mol and 0.91 kcal/mol), 14 (0.86 and 0.87 kcal/mol), and 4 (0.95 and 0.75 kcal/mol). The reactions for which the maximum deviation is observed are all pericyclic reactions, which agrees with the recent report by Sandler *et al.* who showed that DLPNO error is higher for dispersion-dominated and Diels-Alder BHs, with errors that can be as high as 1.2 kcal/mol.[31] The behavior of the errors in Table III confirm the relative difficulty of the DLPNO approximation in modeling large dispersion-dominated systems: All bimolecular BHs are overestimated, and the error for the forward and reverse reactions is approximately the same, indicating that the TSs are predicted to be too unstable by DLPNO. Based on these observations and the fact that the reaction subset in Table II contains the smallest molecules in BH9, we expect the 0.43 kcal/mol to be an overly optimistic error bar. An average error from the DLPNO approximation of around 0.5–1 kcal/mol for the reference data in BH9 is probably a more realistic estimate.

On the grounds of the preceding analysis, it is clear that the DLPNO approximation is the main contributing factor to the error in the BH9 reference data. Since basis set incompleteness is not the leading contribution to the error, our basis set extrapolation approach is justified.[170] Our analysis also shows that the estimated error is low enough to benchmark density functional approximations, which have typical errors in the range of a few kcal/mol[22] (see below). The reference data can be revised in the future as more powerful computers and better algorithms become available.

3.2 Assessment of density functional theory methods

We now proceed to assess a few density functionals that are popular in mechanistic studies with the BH9 set. Our objectives are: i) evaluate whether the increased number of reactions and the larger molecules in the BH9 offer a picture of the performance of these functionals for thermochemistry and kinetics that is different from previous studies,[21,22,158,160] ii) analyze the errors in the BH and RE calculations as a function of reaction type, and iii) benchmark the available XDM-corrected density functionals regarding their ability to calculate REs and BHs, something that has been done previously only with a very limited set of reactions.[160] We expect the inclusion of XDM dispersion to have a similar effect to D3, which has been extensively studied.[22,158]

Tables IV and V show the mean absolute error (MAE) and mean error (ME) of the selected functionals for REs and BHs, grouped by type. The overall RE and BH MAEs are shown graphically in Figure 1. In agreement with previous studies,[22,158,160] the performance of hybrid and range-separated hybrid functionals is, in general, much better than that of GGA functionals. Also, Figure 1 shows that there is a degree of positive correlation between RE and BH average errors, indicating that functionals that perform well for REs tend to work for BHs as well. The best-performing functionals are ωB97XD, M05-2X, and M06-2X with MAEs for both BHs and REs between 2 and 3 kcal/mol. The good performance of these functionals (or variants of ωB97X in combination with other dispersion corrections) has been noted in previous works.[21,22,53] Close in performance, but with slightly higher MAEs are PBE0-XDM (MAE(RE) = 2.74 kcal/mol; MAE(BH) = 2.85 kcal/mol) and CAM-B3LYP-XDM (MAE(RE) = 3.14 kcal/mol; MAE(BH) = 2.37 kcal/mol). B3LYP performs relatively poorly both in RE and BH, and so does its XDM-corrected version with average errors slightly over 4 kcal/mol.

Table IV. Average errors in the BH9 reaction energies using various density functionals.[a]

Functional		I	II	III	IV	V	VI	VII	VIII	IX	Total
BLYP	MAE	8.42	22.27	8.75	3.17	2.35	11.70	2.82	4.01	10.84	11.27
	ME	8.42	20.96	-1.76	-1.24	0.01	10.59	0.87	-0.78	10.84	8.47
BLYP-XDM	MAE	5.27	13.87	7.77	2.77	2.24	3.92	2.70	3.50	4.65	7.15
	ME	5.24	12.51	-1.09	-1.17	-0.02	3.11	0.64	-0.87	4.56	4.61
PBE	MAE	3.34	10.52	8.71	3.24	2.49	5.79	3.00	4.04	3.67	6.22
	ME	1.56	9.36	-1.95	-1.02	1.52	4.78	0.50	-1.29	2.44	3.32
PBE-XDM	MAE	2.68	7.59	8.36	3.12	2.48	2.26	3.00	4.00	3.14	4.87
	ME	0.68	6.36	-1.72	-1.01	1.49	1.36	0.44	-1.43	0.02	1.90
TPSS	MAE	4.10	13.81	8.34	3.14	2.45	8.22	2.48	3.72	5.57	7.54
	ME	2.96	12.84	-1.56	-0.88	1.24	7.12	0.66	-1.24	5.49	4.95
TPSS-XDM	MAE	2.82	8.87	7.73	2.94	2.51	3.01	2.44	3.13	2.87	5.19
	ME	1.26	7.79	-1.16	-0.85	1.24	2.10	0.54	-1.37	1.64	2.62
revTPSS	MAE	3.32	10.91	7.82	3.37	2.82	7.46	2.12	3.76	3.95	6.43
	ME	1.22	9.92	-1.20	-0.76	1.58	6.48	0.64	-1.36	3.76	3.80
B3LYP	MAE	5.63	15.65	7.12	2.40	2.56	8.91	1.81	2.77	8.02	8.18
	ME	5.62	14.75	-0.47	-0.91	-1.67	7.94	0.61	-0.28	8.02	5.90
B3LYP-XDM	MAE	3.69	9.85	6.37	2.06	2.45	2.78	1.76	2.16	3.68	5.26
	ME	3.62	8.89	0.00	-0.87	-1.69	2.20	0.47	-0.44	3.49	3.20
PBE0	MAE	2.69	5.52	6.25	2.19	1.73	4.14	1.77	2.57	2.12	3.78
	ME	-0.81	3.91	-1.05	-0.46	-0.50	3.18	0.23	-0.59	0.92	1.18
PBE0-XDM	MAE	2.40	3.18	5.89	2.02	1.63	1.39	1.77	2.22	2.49	2.74
	ME	-1.67	0.95	-0.83	-0.46	-0.52	-0.22	0.17	-0.72	-1.48	-0.22
BH&HLYP	MAE	3.46	9.42	4.80	1.60	4.11	6.31	0.87	2.07	5.44	5.37
	ME	3.32	8.78	0.14	-0.31	-3.02	5.42	0.45	0.66	5.42	3.53
BH&HLYP-XDM	MAE	2.22	5.36	4.33	1.33	3.97	2.39	0.86	1.43	2.39	3.36
	ME	2.01	4.58	0.46	-0.32	-3.01	0.95	0.36	0.53	2.10	1.57
M05-2X	MAE	1.61	3.12	4.55	1.26	2.54	1.73	0.92	1.83	1.69	2.39
	ME	0.28	2.30	1.10	-0.43	-1.65	0.89	-0.08	0.03	0.75	0.72
M06-2X	MAE	2.31	4.00	4.72	1.56	1.47	2.36	0.91	1.84	1.71	2.76
	ME	0.51	3.71	0.80	-0.63	-0.58	1.76	-0.14	-0.56	1.36	1.30
CAM-B3LYP	MAE	2.98	8.37	5.29	1.62	2.62	5.89	1.08	1.82	4.67	4.82
	ME	2.72	7.83	-0.29	-0.52	-1.60	5.10	0.57	0.07	4.65	3.13
CAM-B3LYP-XDM	MAE	2.10	5.03	4.87	1.43	2.57	2.11	1.08	1.30	2.14	3.14
	ME	1.71	4.42	-0.02	-0.51	-1.64	1.19	0.50	-0.12	1.80	1.51
LC-ωPBE	MAE	4.28	4.38	5.90	1.66	1.73	2.98	1.19	1.82	1.72	3.30
	ME	-3.68	-2.57	0.11	-0.18	1.00	1.68	0.76	0.20	0.01	-0.97
LC-ωPBE-XDM	MAE	4.83	6.84	5.54	1.63	1.69	2.99	1.18	1.50	2.62	4.13
	ME	-4.83	-6.18	0.39	-0.16	1.13	-2.04	0.69	0.21	-2.57	-2.62
ωB97XD	MAE	1.76	3.03	5.27	1.71	1.32	1.94	1.34	1.19	1.69	2.42
	ME	0.08	2.04	-0.08	-0.48	-0.11	-0.42	0.49	-0.22	1.24	0.57

[a]) units are kcal/mol, MAE = mean absolute error, ME = mean error. The roman numerals represent the reaction types in Table II.

Table V. Average errors in the BH9 barrier heights using various density functionals.[a]

Functional		I	II	III	IV	V	VI	VII	VIII	IX	Total
BLYP	MAE	5.01	12.19	7.54	4.87	5.05	7.35	3.40	4.00	5.73	7.62
	ME	-3.77	-5.94	-6.46	-1.97	0.29	2.74	-1.56	-2.09	-3.45	-3.33
BLYP-XDM	MAE	5.05	10.26	11.86	8.51	10.67	4.45	3.62	8.72	6.16	8.66
	ME	-5.05	-10.07	-11.86	-8.49	-10.42	-3.06	-3.53	-8.72	-5.89	-8.44
PBE	MAE	3.85	7.98	8.59	7.01	7.59	3.99	5.51	4.04	4.77	6.68
	ME	-3.69	-6.55	-8.48	-6.62	-6.57	-1.46	-4.86	-3.27	-4.51	-5.78
PBE-XDM	MAE	4.18	8.20	10.99	9.63	12.04	4.39	5.65	6.32	5.57	8.12
	ME	-4.18	-8.16	-10.99	-9.60	-11.96	-4.29	-5.64	-6.29	-5.44	-8.07
TPSS	MAE	3.89	9.04	6.88	4.84	4.78	5.23	3.84	4.04	4.44	6.19
	ME	-3.52	-6.41	-6.38	-3.08	-1.81	-0.10	-2.63	-3.13	-3.85	-4.17
TPSS-XDM	MAE	4.33	9.07	10.04	7.56	9.66	4.30	3.97	7.57	5.47	7.66
	ME	-4.33	-9.05	-10.04	-7.52	-9.47	-4.13	-3.85	-7.57	-5.36	-7.61
revTPSS	MAE	3.86	8.12	6.99	4.53	4.43	5.03	3.14	4.08	4.19	5.78
	ME	-3.65	-6.54	-6.66	-2.92	-2.06	-1.13	-1.83	-3.44	-3.85	-4.31
B3LYP	MAE	2.98	8.08	4.27	3.66	5.69	5.88	1.75	3.58	4.10	5.37
	ME	-1.04	-0.89	-1.08	1.08	4.81	4.23	0.11	2.02	-0.95	0.52
B3LYP-XDM	MAE	2.34	5.51	5.81	4.03	4.05	2.43	1.58	3.09	3.13	4.22
	ME	-1.97	-3.91	-5.17	-3.95	-3.73	-0.33	-1.28	-2.99	-2.68	-3.38
PBE0	MAE	1.46	3.34	3.47	3.08	3.82	2.72	2.76	3.85	1.62	3.00
	ME	-0.28	-0.05	-0.76	-0.95	1.53	1.56	-1.91	2.93	-1.01	-0.05
PBE0-XDM	MAE	1.43	2.54	3.87	3.98	4.09	1.55	2.76	1.52	2.13	2.85
	ME	-0.77	-1.66	-3.25	-3.91	-3.89	-1.28	-2.69	-0.06	-1.93	-2.33
BH&HLYP	MAE	2.55	7.27	7.54	6.82	14.49	7.21	3.36	7.03	3.39	7.05
	ME	2.25	6.68	6.69	6.81	14.49	7.18	3.36	6.98	3.07	6.73
BH&HLYP-XDM	MAE	1.81	4.92	4.43	2.99	7.50	3.71	2.30	3.22	1.96	4.01
	ME	1.55	4.41	3.45	2.85	7.48	3.52	2.30	3.06	1.77	3.67
M05-2X	MAE	0.96	1.94	2.90	1.56	6.30	1.72	1.26	2.97	1.06	2.21
	ME	0.03	0.15	1.38	0.36	5.70	0.53	-0.32	2.25	-0.57	0.86
M06-2X	MAE	1.61	2.39	2.96	1.36	4.99	1.66	1.11	3.14	1.13	2.27
	ME	1.06	0.92	2.03	-0.30	4.44	0.37	-0.22	2.76	-0.28	1.05
CAM-B3LYP	MAE	1.83	5.06	4.18	3.20	9.34	5.21	1.21	4.89	2.41	4.43
	ME	1.19	3.86	2.28	2.47	9.05	5.05	0.16	4.65	1.29	3.52
CAM-B3LYP-XDM	MAE	1.27	3.16	2.55	1.68	3.50	2.26	1.07	2.13	1.19	2.37
	ME	0.63	2.03	-0.46	-0.86	3.13	1.85	-0.70	1.31	0.25	0.96
LC-ωPBE	MAE	3.70	9.15	8.47	4.82	13.79	5.38	1.40	10.38	3.70	7.33
	ME	3.49	9.12	8.31	4.66	13.37	5.18	0.13	10.38	3.63	7.16
LC-ωPBE-XDM	MAE	3.16	7.15	5.59	1.52	7.35	2.60	1.22	6.87	2.84	4.72
	ME	2.88	7.13	5.32	1.12	6.86	2.09	-0.89	6.87	2.46	4.42
ωB97XD	MAE	1.13	2.66	2.97	1.73	2.17	1.58	0.90	2.99	1.21	2.10
	ME	0.65	2.26	1.91	-1.20	1.46	0.80	0.04	2.79	0.58	1.04

[a] units are kcal/mol, MAE = mean absolute error, ME = mean error. The roman numerals represent the reaction types in Table II.

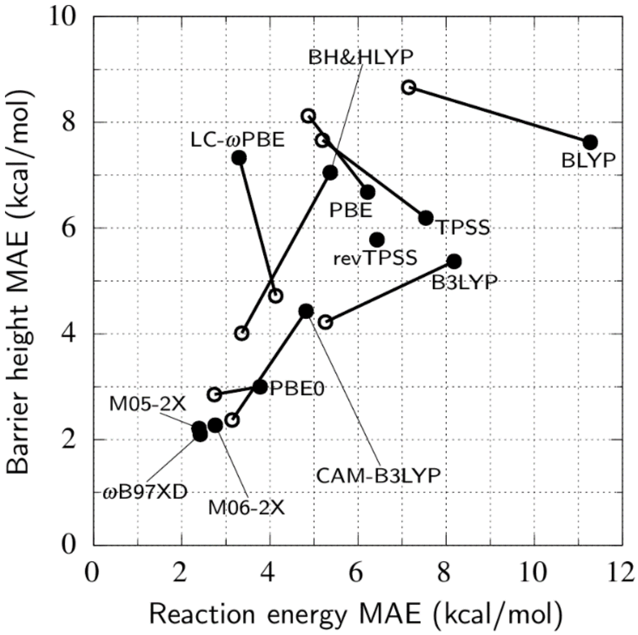


Figure 1. Barrier height vs. reaction energy mean absolute errors (MAE) for the chosen functionals. Open symbols represent the XDM-corrected version of each functional.

The good performance of M05-2X and M06-2X is interesting. It is known that these functionals underestimate non-covalent interaction energies at long range.[181,182] Since the importance of long-range dispersion increases with molecular size,[183] one would have expected a degradation in the performance of these functionals for the BH9 relative to previous studies of REs and BHs involving smaller molecules.[22,181] However, this does not seem to be the case, and our average errors are similar to those reported by Mardirossian *et al.*[181] and Goerigk *et al.*[22] Because TSs of addition reactions are larger than either the reactant or product molecules, an underestimation of non-covalent binding would lead to an erroneously unstable TS and an overestimation of the BHs. This seems to be the case for M05-2X and M06-2X, as indicated by the MEs in Table V. However, the average bias is only 0.72 (M05-2X) and 1.30 kcal/mol (M06-2X) for REs and 0.86 (M05-2X) and 1.05 kcal/mol (M06-2X) for BHs, suggesting that capturing the correct asymptotic dependence of the dispersion contribution seems not to be as important for the calculation of REs and BHs as previously argued.[22] This point is reinforced by the fact that the performance of M05-2X and M06-2X in the GMTKN database is only marginally improved by their combination with the D3 dispersion correction.[22,158]

Compared to our previous analysis of the performance of XDM-corrected functionals for REs and BHs,[160] the advantages of a more complete benchmark set are very evident. In our previous work, LC-ωPBE-XDM (MAE = 1.43 kcal/mol) and BH&HLYP-XDM (MAE = 2.38 kcal/mol) were the best-performing functionals for BHs.[160] This is in stark contrast with the results in Table V and Figure 1, where the MAEs of these functionals rise to 4.72 and 4.01 kcal/mol, respectively. The cause of this disagreement is very likely the limited size of the benchmark set used in our previous work,[160] the small size of the molecular species, and the fact that it contained only hydrogen atom transfer reactions. Still, the results in Table IV, Table V, and Figure 1 are encouraging and suggest that expanding the list of functionals with which XDM has been combined could increase the applicability of the method to chemical problems other than modeling non-covalent interactions.

As expected, the effect of including the XDM dispersion energy agrees, in general terms, with previous reports in the literature using D3.[22,30,158,184,185] The inclusion of XDM has a noticeable impact on REs and BHs. Uncorrected GGAs severely overestimate REs, with an MAE that can be as high as 11.27 kcal/mol (BLYP). The overestimation is less pronounced for uncorrected hybrid and range-separated hybrid functionals. The inclusion of XDM dispersion partially corrects the overestimation of the REs and reduces the MAEs by several kcal/mol in general, except in the case of LC-ωPBE. These observations can be explained by the fact that the overall RE error is dominated by addition reactions, where the product molecule is the combination of both reactants. Functionals without dispersion underestimate the stability of the addition products, resulting in erroneously high REs.

In the case of BHs, Figure 1 and Table V show that the MAEs for the uncorrected functionals are in the range 4–8 kcal/mol. In particular, all GGA and meta-GGA functionals severely underestimate the BHs. This is explained by delocalization error,[28] the tendency of approximate density functionals to over stabilize delocalized molecules. The TS are, in general, more delocalized than reactants and products. Consequently, they are spuriously stabilized, resulting in an erroneously low BH.[28] Delocalization error severely affects GGAs, while admixture of exact exchange in global and range-separated hybrids mitigates, but does not eliminate, this problem. The effect of including the XDM dispersion energy on BHs can be understood as well. The dispersion stabilization increases with the size of the molecule, so including the dispersion energy always leads to lower BHs. For GGAs functionals, which spuriously underestimate BHs, inclusion of XDM results in an increased MAE. For hybrid and range-separated hybrid functionals, which do not suffer as much from delocalization error, the use of XDM decreases the MAEs. This is consistent with previous analysis in the literature.[28,158,160]

We now analyze the performance of the chosen functionals on the various reaction types of the BH9 set using the data in Tables IV and V. For a few representative functionals, the RE and BH MAEs as a function of reaction type are shown in Figure 2. In the case of the REs, there are large differences between reaction types regarding the performance of various functionals and the effect of dispersion. Reaction types IV (hydrogen atom transfer), V (hydride transfer), VII (proton transfer), and VIII (nucleophilic substitution) seem to be modellable with approximately the same error by all uncorrected and dispersion-corrected functionals, in the range 2–4 kcal/mol. However, hybrid and range-separated hybrid functionals are, again, slightly better than GGAs. Reactions I (mostly radical rearrangements) and III (halogen atom transfer) show higher errors (up to around 8 kcal/mol for BLYP), are better modeled by hybrid or range-separated functionals, and the inclusion of dispersion corrections has a relatively minor impact. Reaction types VI (B- and Si-containing reactions) and IX (nucleophilic addition) show similar or larger errors than I and III and hybrid and range-separated hybrid functionals outperform GGAs. However, in this case, the inclusion of dispersion interactions improves the functional performance by several kcal/mol. These observations are easily explained by the fact that types VI and IX comprise addition reactions, while the other reaction types mentioned are rearrangements or atom transfer reactions. Since dispersion interactions stabilize larger molecules, their inclusion alleviates the overestimation of the REs in these reactions by the uncorrected functionals, as mentioned above. (Note that most REs for addition reactions are negative. Uncorrected functionals yield overestimated REs in general; their REs are above this negative reference value, but smaller in magnitude.) Lastly, the pericyclic reactions (II) show the highest errors, possibly due to the effect of varying delocalization between reactants and products, and benefit from dispersion for the same reason as VI and IX, since most of the members of this category are addition reactions.

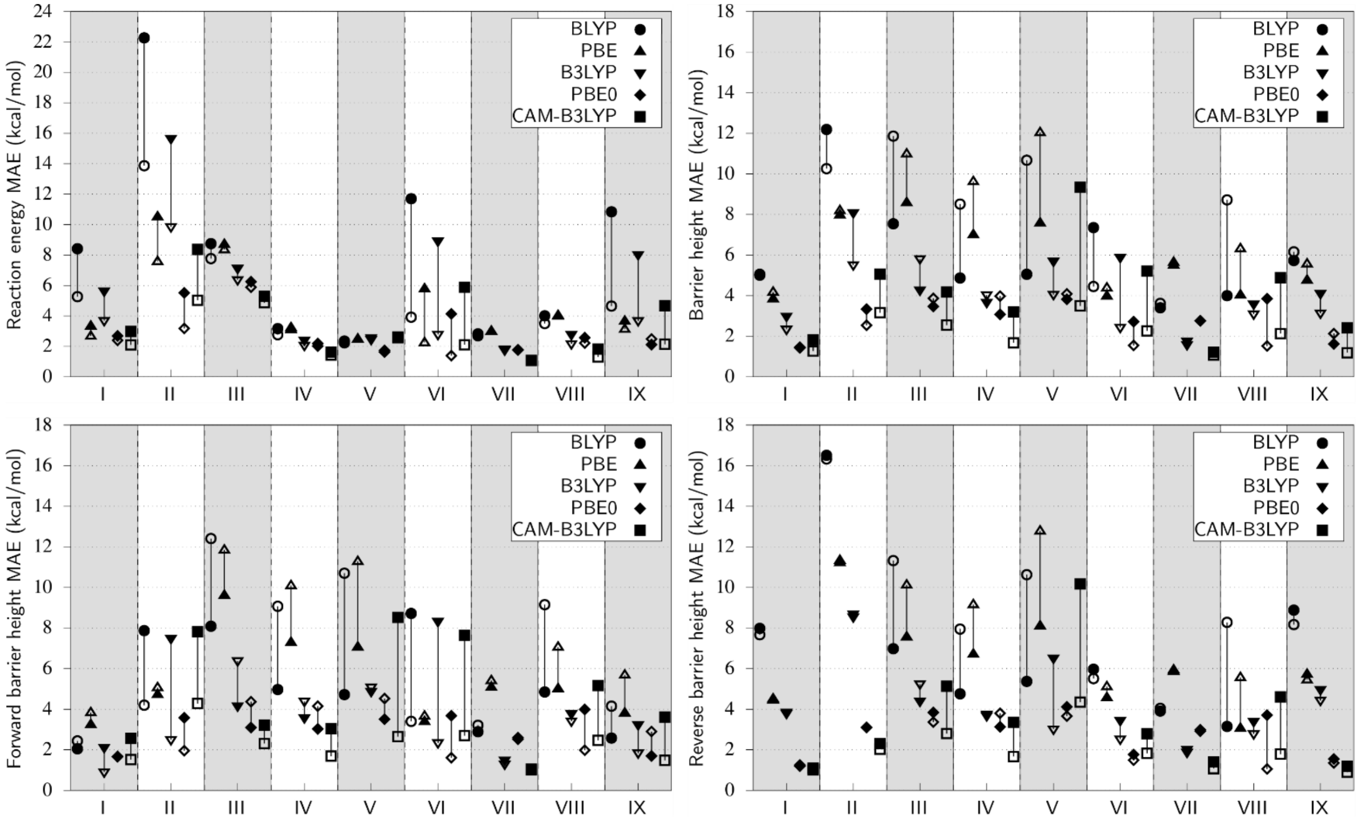


Figure 2. Reaction energy (top left), both forward and reverse barrier height (top right), forward reaction barrier height (bottom left), and reverse reaction barrier height (bottom right) mean absolute errors (MAEs) as a function of reaction type and density functional (using the Def2-QZVPP basis set). Open symbols represent the XDM-corrected version of each functional.

Figure 2 and Table V show that the MAEs for BHs are higher than for REs, and that the inclusion of dispersion has comparatively more impact. BHs are more accurately represented by hybrid functionals, particularly if they are dispersion corrected, than by GGA functionals for all reaction types. For types I (radical rearrangement) and VII (proton transfer), the effect of including dispersion is minimal, and the accuracy is entirely controlled by the base functional, with hybrid and range-separated hybrid functionals showing much better performance. In reactions II (pericyclic) and VI (B- and Si-containing reactions), including dispersion interactions either has no effect or is beneficial, regardless of the functional type. Figure 2 also shows the MAE for the forward and reverse BHs separately. Reactions II and VI are particular in that the effect of dispersion is very noticeable in the forward reaction BHs, but it is not for the reverse reaction BHs. This is reasonable because both categories comprise addition reactions. The dispersion stabilization of the product is essentially the same as the TS, but higher than for the reactant molecules. For the rest of the reactions, the inclusion of dispersion increases the MAE of the GGA functionals and decreases (in general) the MAEs of hybrid and range-separated hybrid functionals, for the reasons stated above.

The fact that the inclusion of dispersion interactions decreases the MAE for (forward) BHs in pericyclic reactions (II) is slightly surprising in light of our previous discussion regarding delocalization error. Given the delocalized nature of the TS in pericyclic reactions, we expected a severe underestimation of the forward BH by GGAs and a subsequent increase in the MAE upon application of XDM. We can interpret this by noting that the reacting molecules are larger in the pericyclic reactions than in other reactions of the BH9 set, which suggests that non-covalent interactions have a comparatively more important role in the stabilization of the TS than electronic delocalization. Omitting dispersion interactions from the functional destabilizes the TS more than the spurious stabilizing effect from delocalization error, and therefore the inclusion of XDM is beneficial.

In summary, our analysis shows that, in agreement with previous studies,[21,22] the performance of density functional approximations improves with sophistication: GGA and meta-GGA functionals are not usable in general for RE and BH prediction while hybrid and range-separated hybrid functionals offer smaller average errors. The best performers among the functionals studied are the ωB97XD and the M05-2X and M06-2X Minnesota functionals, which agrees with previous works,[21,22] where these functionals, or variants of them, were shown to be among the best functionals available for the calculation of REs and BHs (except for double-hybrids). However, our results also show that functional performance depends strongly on the type of reaction studied and, comparison with the average errors reported in the literature[21,22] shows that errors in the calculation of BHs and REs are higher for larger systems.

3.3 Assessment of basis set incompleteness potentials

One of our objectives in the construction of the BH9 set is to provide training data for the development of atom-centered potentials (ACPs).[55–57,186] ACPs are one-electron potentials that are designed to correct for the shortcomings of the DFT method to which they are applied. One particular flavor of ACPs are the basis set incompleteness potentials[55,56] (BSIPs) whose purpose is to minimize the basis set incompleteness error (BSIE) that originates from using small or minimal basis sets in DFT calculations. The application of BSIPs allows computing molecular properties with a quality similar to a complete basis set but at a much reduced computational cost.

The development of ACPs requires a relatively large training set of molecular properties. BSIPs, in particular, are constructed by minimizing the deviation between the BSIP-corrected small-basis-set values and the complete-basis-set values for a number of molecular properties. Since both the approximate and the reference molecular properties are calculated using the same functional and BSIE is mostly functional-independent, this ensures that BSIPs are mostly transferable between functionals, and are tied only to the basis set for which they were developed.[55]

The training set for the recently developed BSIPs contained only 316 REs and 102 BHs, out of a total of 9,372 molecular properties.[56] Therefore, it is interesting to examine whether BSIPs decrease BSIE in the calculation of REs and BHs by applying them to the BH9 set. For this test, we used the BSIPs from our previous work.[56] Of the 15 basis sets for which BSIPs were developed, only six had thermochemical data in their training set (6-31G*, 6-31+G*, 6-31+G**, Def2-SV(P), Def2-SVP, and pc-1), so we restrict our analysis to these basis sets. We chose PBE0-XDM as the base functional for this analysis in order to check the transferability of BSIPs across functionals. (These BSIPs were developed using B3LYP.[56])

The MAEs for the REs and BHs in the BH9 set using BSIP-corrected and uncorrected PB0-XDM in combination with the aforementioned basis sets are shown in Table VI. The table shows two sets of MAEs with respect to different reference data: the DLPNO-CCSD(T)/CBS values ("Ref.") and our complete-basis-set PBE0-XDM estimate using the Def2-QZVPP basis set ("CBS"). Since the objective of BSIP development is to minimize BSIE, comparison with the CBS results is the purest measure of performance.

The uncorrected MAEs ("Bare") in Table VI show that the magnitude of the BSIE on average is between 1.4 kcal/mol and 4 kcal/mol, depending on the basis set, and that there are no significant differences between BHs and REs regarding BSIE. When BSIPs are applied, the MAEs with respect to the CBS values decreases in all cases, by up to 2 kcal/mol, bringing the results to a reasonably close agreement with the Def2-QZVPP results. The performance of BSIPs is better for the larger basis sets, 6-31+G* and 6-31+G**, where the MAEs with respect to the CBS reference are lower than 1 kcal/mol for both BHs and REs. There seems to be no salient differences between the effect of BSIPs on REs and BHs. These results are encouraging because of the aforementioned sparsity of thermochemical and kinetic data in the training set, which suggests that BSIPs have robust performance for systems significantly different from those in the training set. Also, because a functional different from PBE0-XDM was used in their development, our results suggest a strong transferability of BSIPs across functionals.

Table VI. Mean absolute errors (MAE) for the BH9 barrier heights and reaction energies of PBE0-XDM with several BSIP-corrected and uncorrected basis sets.[a]

		Reaction energies		**Barrier heights**	
		CBS[b]	**Ref.**[c]	**CBS**[b]	**Ref.**[c]
6-31G*	Bare	3.59	3.97	2.46	4.27
	BSIP	1.91	3.46	1.62	3.26
6-31+G*	Bare	2.49	3.01	1.48	3.30
	BSIP	0.89	2.65	0.75	2.83
6-31+G**	Bare	1.81	2.66	1.42	3.48
	BSIP	0.64	2.65	0.63	2.81
Def2-SV(P)	Bare	4.12	4.65	3.11	4.96
	BSIP	1.95	3.64	1.80	2.91
Def2-SVP	Bare	3.12	3.97	2.90	4.87
	BSIP	2.19	3.97	1.74	3.06
pc-1	Bare	2.42	2.86	2.60	4.80
	BSIP	1.41	3.17	1.32	3.18
Def2-QZVPP			2.74		2.85

[a]) units are kcal/mol. The statistics correspond to the whole BH9 set. [b]) CBS = MAEs calculated with respect to our complete-basis-set estimate (Def2-QZVPP). [c]) Ref. = MAEs calculated with respect to the DLPNO-CCSD(T)/CBS reference data for the BH9.

The MAEs between the small-basis-set BSIP-corrected and uncorrected PBE0-XDM results and the DLPNO-CCSD(T)/CBS data are also shown in Table VI ("Ref." column). In this case, the particular MAE values result from a combination of two errors: the uncorrected BSIE and the errors from the PBE0-XDM functional itself. Application of BSIPs reduces the MAEs in general to values that are close to the MAE of PBE0-XDM/Def2-QZVPP, particularly for the BHs, for which the BSIP-corrected MAEs are at most

0.41 kcal/mol above the Def2-QZVPP MAE. However, in some cases the MAE is unaffected or increases slightly due to favorable error cancellation in the uncorrected results.

4. Conclusions

In this article we introduce the BH9 set, an extensive and diverse benchmark dataset for reaction energies (REs) and barrier heights (BHs) in organic and bio-organic reactions. The BH9 set comprises 449 diverse reactions (449 REs and 898 BHs) involving relatively large molecular species (up to 71 atoms), similar to those found in thermochemical and mechanistic studies. The molecular species in BH9 comprise main-group elements, particularly those typically found in organic and bio-organic chemistry (H, C, N, O, F, P, S, and Cl) plus B and Si.

The computational level for the BH9 reference data is DLPNO-CCSD(T) combined with a focal-point approach in order to minimize errors from basis set incompleteness. We used a small subset of the BH9 composed of small molecular species to evaluate the errors introduced by our approximations and to estimate an error bar for the BH9 reference data. The DLPNO approximation is the main source of error, in comparison with basis set incompleteness. We estimate that the overall accuracy of the benchmark is in the vicinity of 1 kcal/mol or better.

The newly created BH9 was applied in two ways. First, we benchmarked a few popular density functionals used in the literature for calculating REs and BHs, as well as some XDM-corrected functionals to evaluate the effect of dispersion interactions on REs and BHs. In general, hybrid and range-separated hybrid functionals perform much better than GGA functionals. The two Minnesota functionals M05-2X and M06-2X and the ωB97XD functional had the lowest mean absolute errors (MAEs), between 2 and 3 kcal/mol for both BHs and REs. The XDM-corrected functionals PBE0-XDM and CAM-B3LYP-XDM closely followed these functionals in terms of performance, with MAEs not above 3 kcal/mol. We also verified that delocalization error is a major contribution to the BH and RE errors. However, for the reactions involving large molecular species (e.g., some pericyclic reactions), the incorrect treatment of dispersion seems to outweigh delocalization error for non-dispersion-corrected functionals.

Lastly, we applied the BH9 set to analyze the performance of our basis set incompleteness potentials (BSIPs) for REs and BHs in combination with a few double-ζ basis sets and the PBE0-XDM functional. We found that, despite the fact that thermochemical and kinetic data were only a small part of their training set and that they were developed using a different functional (B3LYP), BSIPs performed excellently, reducing the discrepancy between the double-ζ and the complete-basis-set results by a factor of around 1.5

to 2. For BHs, the BSIP-corrected double-ζ MAEs were at most 0.41 kcal/mol higher than the Def2-QZVPP MAE, and the calculations were immensely less expensive. This confirms BSIPs are a robust way of minimizing basis set incompleteness from finite basis sets.

To our knowledge, BH9 is the most comprehensive BH and RE benchmark set to date. We hope that it will be useful to assess and develop new methods for thermochemical and kinetic work.

References

(1) Laidler, K. Chemical Kinetics; Harper & Row, 1987.

(2) Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. Current status of transition-state theory. *J. Phys. Chem.* **1996**, *100*, 12771–12800.

(3) Klippenstein, S. J.; Pande, V. S.; Truhlar, D. G. Chemical kinetics and mechanisms of complex systems: a perspective on recent theoretical advances. *J. Am. Chem. Soc.* **2014**, *136*, 528–546.

(4) Bachrach, S. M. Challenges in computational organic chemistry. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 482–487.

(5) Houk, K.; Liu, F. Holy grails for computational organic chemistry and biochemistry. *Acc. Chem. Res.* **2017**, *50*, 539–543.

(6) van der Kamp, M. W.; Mulholland, A. J. Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. *Biochemistry* **2013**, *52*, 2708–2728.

(7) Ahmadi, S.; Barrios Herrera, L.; Chehelamirani, M.; Hostaˇs, J.; Jalife, S.; Salahub, D. R. Multiscale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: A tutorial review. *Int. J. Quantum Chem*. **2018**, *118*, e25558.

(8) Himo, F. Recent trends in quantum chemical modeling of enzymatic reactions. J. Am. Chem. Soc. 2017, 139, 6780–6786.

(9) Quesne, M. G.; Borowski, T.; de Visser, S. P. Quantum mechanics/molecular mechanics modeling of enzymatic processes: Caveats and breakthroughs. *Chem. Eur. J.* **2016**, *22*, 2562–2581.

(10) Engkvist, O.; Norrby, P.-O.; Selmi, N.; Lam, Y.-h.; Peng, Z.; Sherer, E. C.; Amberg, W.; Erhard, T.; Smyth, L. A. Computational prediction of chemical reactions: current status and outlook. *Drug Discov. Today* **2018**, *23*, 1203–1218.

(11) Lam, Y.-h.; Abramov, Y.; Ananthula, R. S.; Elward, J. M.; Hilden, L. R.; Nilsson Lill, S. O.; Norrby, P.-O.; Ramirez, A.; Sherer, E. C.; Mustakis, J., et al. Applications of Quantum Chemistry in Pharmaceutical Process Development: Current State and Opportunities. *Org. Process Res. Dev.* **2020**, *24*, 1496–1507.

(12) Peterson, K. A.; Feller, D.; Dixon, D. A. Chemical accuracy in ab initio thermochemistry and spectroscopy: current strategies and future challenges. *Theor. Chem. Acc.* **2012**, *131*, 1–20.

(13) Karton, A.; Rabinovich, E.; Martin, J. M.; Ruscic, B. W4 theory for computational thermochemistry: In pursuit of confident sub-kJ/mol predictions. *J. Chem. Phys.* **2006**, *125*, 144108.

(14) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **2007**, *126*, 084108.

(15) Harding, M. E.; V´azquez, J.; Ruscic, B.; Wilson, A. K.; Gauss, J.; Stanton, J. F. High-accuracy extrapolated ab initio thermochemistry. III. Additional improvements and overview. *J. Chem. Phys.* **2008**, *128*, 114111.

(16) Karton, A. A computational chemist's guide to accurate thermochemistry for organic molecules. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2016**, *6*, 292–310.

(17) Karton, A.; Sylvetsky, N.; Martin, J. M. W4-17: A diverse and high-confidence dataset of atomization energies for benchmarking high-level electronic structure methods. *J. Comput. Chem.* **2017**, *38*, 2063–2075.

(18) Chan, B. How to computationally calculate thermochemical properties objectively, accurately, and as economically as possible. *Pure Appl. Chem.* **2017**, *89*, 699–713.

(19) Bistoni, G.; Polyak, I.; Sparta, M.; Thiel, W.; Neese, F. Toward accurate QM/MM reaction barriers with large QM regions using domain based pair natural orbital coupled cluster theory. *J. Chem. Theory Comput.* **2018**, *14*, 3524–3531.

(20) Deglmann, P.; Sch¨afer, A.; Lennartz, C. Application of quantum calculations in the chemical industry—An overview. *Int. J. Quantum Chem.* **2015**, *115*, 107–136.

(21) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.

(22) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184–32215.

(23) Goerigk, L.; Mehta, N. A trip to the density functional theory zoo: warnings and recommendations for the user. *Aust. J. Chem.* **2019**, *72*, 563–573.

(24) Morgante, P.; Peverati, R. ACCDB: A collection of chemistry databases for broad computational purposes. *J. Comput. Chem.* **2019**, *40*, 839–848.

(25) Yu, H. S.; Zhang, W.; Verma, P.; He, X.; Truhlar, D. G. Nonseparable exchange–correlation functional for molecules, including homogeneous catalysis involving transition metals. *Phys. Chem. Chem. Phys.* **2015**, *17* .

(26) Yu, H. S.; He, X.; Truhlar, D. G. MN15-L: A new local exchange-correlation functional for Kohn–Sham density functional theory with broad accuracy for atoms, molecules, and solids. *J. Chem. Theory Comput.* **2016**, *12*, 1280–1293.

(27) Haoyu, S. Y.; He, X.; Li, S. L.; Truhlar, D. G. MN15: A Kohn–Sham global-hybrid exchange–correlation density functional with broad accuracy for multi-reference and single-reference systems and noncovalent interactions. *Chem. Sci.* **2016**, *7*, 5032–5051.

(28) Johnson, E. R.; Mori-S´anchez, P.; Cohen, A. J.; Yang, W. Delocalization errors in density functionals and implications for main-group thermochemistry. *J. Chem. Phys.* **2008**, *129*, 204112.

(29) Yu, L.-J.; Sarrami, F.; O'Reilly, R. J.; Karton, A. Reaction barrier heights for cycloreversion of heterocyclic rings: An Achilles' heel for DFT and standard ab initio procedures. *Chem. Phys.* **2015**, *458*, 1–8.

(30) Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. Inclusion of dispersion effects significantly improves accuracy of calculated reaction barriers for cytochrome P450 catalyzed reactions. *J. Phys. Chem. Lett.* **2010**, *1*, 3232–3237.

(31) Sandler, I.; Chen, J.; Taylor, M.; Sharma, S.; Ho, J. Accuracy of DLPNO-CCSD (T): Effect of Basis Set and System Size. *J. Phys. Chem. A* **2021**, *125*, 1553–1563.

(32) Wappett, D. A.; Goerigk, L. Toward a quantum-chemical benchmark set for enzymatically catalyzed reactions: important steps and insights. *J. Phys. Chem. A* **2019**, *123*, 7057–7074.

(33) Wappett, D. A.; Goerigk, L. A guide to benchmarking enzymatically catalysed reactions: the importance of accurate reference energies and the chemical environment. *Theor. Chem. Acc.* **2021**, *140*, 1–15.

(34) Zheng, J.; Zhao, Y.; Truhlar, D. G. Representative benchmark suites for barrier heights of diverse reaction types and assessment of electronic structure methods for thermochemical kinetics. *J. Chem. Theory Comput.* **2007**, *3*, 569–582.

(35) Karton, A.; Tarnopolsky, A.; Lam´ere, J.-F.; Schatz, G. C.; Martin, J. M. Highly accurate first-principles benchmark data sets for the parametrization and validation of density functional and other approximate methods. Derivation of a robust, generally applicable, double-hybrid functional for thermochemistry and thermochemical kinetics. *J. Phys. Chem. A* **2008**, *112*, 12868–12886.

(36) Zhao, Y.; Gonz´alez-Garc´ıa, N.; Truhlar, D. G. Benchmark database of barrier heights for heavy atom transfer, nucleophilic substitution, association, and unimolecular reactions and its use to test theoretical methods. *J. Phys. Chem. A* **2005**, *109*, 2012–2018.

(37) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. Multi-coefficient extrapolated density functional theory for thermochemistry and thermochemical kinetics. *Phys. Chem. Chem. Phys.* **2005**, *7*, 43–52.

(38) Chan, B.; Simmie, J. M. Barriometry–an enhanced database of accurate barrier heights for gas-phase reactions. *Phys. Chem. Chem. Phys.* **2018**, *20*, 10732–10740.

(39) Goerigk, L.; Sharma, R. The INV24 test set: how well do quantum-chemical methods describe inversion and racemization barriers? *Can. J. Chem.* **2016**, *94*, 1133–1143.

(40) Karton, A.; Goerigk, L. Accurate reaction barrier heights of pericyclic reactions: Surprisingly large deviations for the CBS-QB3 composite method and their consequences in DFT benchmark studies. *J. Comput. Chem.* **2015**, *36*, 622–632.

(41) Karton, A.; O'Reilly, R. J.; Radom, L. Assessment of theoretical procedures for calculating barrier heights for a diverse set of water-catalyzed proton-transfer reactions. *J. Phys. Chem. A* **2012**, *116*, 4211–4221.

(42) Karton, A.; O'Reilly, R. J.; Chan, B.; Radom, L. Determination of Barrier Heights for Proton Exchange in Small Water, Ammonia, and Hydrogen Fluoride Clusters with G4 (MP2)-Type, MP n, and SCS-MPn Procedures-A Caveat. *J. Chem. Theory Comput.* **2012**, *8*, 3128–3136.

(43) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Sci. Data* **2020**, *7*, 1–8.

(44) von Rudorff, G. F.; Heinen, S. N.; Bragato, M.; von Lilienfeld, O. A. Thousands of reactants and transition states for competing E2 and S2 reactions. *Mach. Learn. Sci. Technol.* **2020**, *1*, 045026.

(45) Kromann, J. C.; Christensen, A. S.; Cui, Q.; Jensen, J. H. Towards a barrier height benchmark set for biologically relevant systems. *PeerJ* **2016**, *4*, e1994.

(46) Riplinger, C.; Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.* **2013**, *138*, 034106.

(47) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *J. Chem. Phys.* **2013**, *139*, 134101.

(48) Liakos, D. G.; Sparta, M.; Kesharwani, M. K.; Martin, J. M.; Neese, F. Exploring the accuracy limits of local pair natural orbital coupled-cluster theory. *J. Chem. Theory Comput.* **2015**, *11*, 1525–1539.

(49) Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F. Sparse maps-A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. *J. Chem. Phys.* **2016**, *144*, 024109.

(50) Saitow, M.; Becker, U.; Riplinger, C.; Valeev, E. F.; Neese, F. A new near-linear scaling, efficient and accurate, open-shell domain-based local pair natural orbital coupled cluster singles and doubles theory. *J. Chem. Phys.* **2017**, *146*, 164105.

(51) Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. Communication: An improved linear scaling perturbative triples correction for the domain based local pair-natural orbital based singles and doubles coupled cluster method (DLPNO-CCSD (T)). *J. Chem. Phys*. **2018**, *148*, 011101.

(52) Paiva, P.; Ramos, M. J.; Fernandes, P. A. Assessing the validity of DLPNO-CCSD(T) in the calculation of activation and reaction energies of ubiquitous enzymatic reactions. *J. Comput. Chem.* **2020**, *41*, 2459–2468.

(53) Iron, M. A.; Janes, T. Evaluating transition metal barrier heights with the latest density functional theory exchange–correlation functionals: the MOBH35 benchmark database. *J. Phys. Chem. A* **2019**, *123*, 3761–3781.

(54) Dohm, S.; Hansen, A.; Steinmetz, M.; Grimme, S.; Checinski, M. P. Comprehensive thermochemical benchmark set of realistic closed-shell metal organic reactions. *J. Chem. Theory Comput.* **2018**, *14*, 2596–2608.

(55) Otero-de-la-Roza, A.; DiLabio, G. A. Transferable atom-centered potentials for the correction of basis set incompleteness errors in density-functional theory. *J. Chem. Theory Comput.* **2017**, *13*, 3505–3524.

(56) Otero-de-la-Roza, A.; DiLabio, G. A. Improved basis-set incompleteness potentials for accurate dft calculations in large systems. *J. Chem. Theory Comput.* **2020**, *16*, 4176–4191.

(57) Prasad, V. K.; Otero-de-la-Roza, A.; DiLabio, G. A. Atom-centered potentials with dispersion-corrected minimal basis set Hartree-Fock: an efficient and accurate computational approach for large molecular systems. *J. Chem. Theory Comput.* **2018**, *14*, 726–738.

(58) Prasad, V. K.; Otero-de La-Roza, A.; DiLabio, G. A. PEPCONF, a diverse data set of peptide conformational energies. *Sci. Data* **2019**, 6, 1–9.

(59) Prasad, V. K.; Khalilian, M. H.; Otero-de-la-Roza, A.; DiLabio, G. A. BSE49, a diverse, high-quality benchmark dataset of separation energies of chemical bonds. *Sci. Data* **2021**, (submitted).

(60) Chan, B. Formulation of Small Test Sets Using Large Test Sets for Efficient Assessment of Quantum Chemistry Methods. *J. Chem. Theory Comput.* **2018**, *14*, 4254–4262.

(61) Ribeiro, A. J. M.; Holliday, G. L.; Furnham, N.; Tyzack, J. D.; Ferris, K.; Thornton, J. M. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res*. **2018**, *46*, D618–D623.

(62) Newcomb, M. In Encyclopedia of Radicals in Chemistry, Biology and Materials; Chatgilialoglu, C., Studer, A., Eds.; Wiley Online Library, 2012.

(63) Mackie, I. D.; DiLabio, G. A. Ring-opening radical clock reactions: many density functionals have difficulty keeping time. *Org. Biomol. Chem.* **2011**, *9*, 3158–3164.

(64) Lucas, M. d. F.; Ramos, M. J. Theoretical study of the suicide inhibition mechanism of the enzyme pyruvate formate lyase by methacrylate. *J. Am. Chem. Soc.* **2005**, *127*, 6902–6909.

(65) Zou, Y.; Xue, X.-S.; Deng, Y.; Smith III, A. B.; Houk, K. Factors Controlling Reactivity in the Hydrogen Atom Transfer and Radical Addition Steps of a Radical Relay Cascade. *Org. Lett.* **2019**, *21*, 5894–5897.

(66) Romero-Silva, A.; Mora-Diez, N.; Alvarez-Idaboy, J. R. Theoretical study of the reactivity and selectivity of various free radicals with cysteine residues. *ACS Omega* **2018**, *3*, 16519–16528.

(67) Hayden, A. E.; Paton, R. S.; Becker, J.; Lim, Y. H.; Nicolaou, K.; Houk, K. Origins of regioselectivity of Diels- Alder reactions for the synthesis of bisanthraquinone antibiotic BE-43472B. *J. Org. Chem.* **2010**, *75*, 922–928.

(68) Lan, Y.; Danheiser, R. L.; Houk, K. Why Nature Eschews the Concerted (2+ 2+ 2) Cycloaddition of a Nonconjugated Cyanodiyne. Computational Study of a Pyridine Synthesis Involving an Ene–Diels–Alder–Bimolecular Hydrogen-Transfer Mechanism. *J. Org. Chem.* **2012**, *77*, 1533–1538.

(69) Zheng, Y.; Thiel, W. Computational insights into an enzyme-catalyzed (4+ 2) cycloaddition. *J. Org. Chem.* **2017**, *82*, 13563–13571.

(70) Sato, M.; Yagishita, F.; Mino, T.; Uchiyama, N.; Patel, A.; Chooi, Y.-H.; Goda, Y.; Xu, W.; Noguchi, H.; Yamamoto, T., et al. Involvement of lipocalin-like CghA in decalin-forming stereoselective intramolecular (4+ 2) cycloaddition. *ChemBioChem* **2015**, *16*, 2294.

(71) Zhang, B.; Wang, K. B.; Wang, W.; Wang, X.; Liu, F.; Zhu, J.; Shi, J.; Li, L. Y.; Han, H.; Xu, K., et al. Enzyme-catalysed (6+ 4) cycloadditions in the biosynthesis of natural products. *Nature* **2019**, *568*, 122–126.

(72) Maiga-Wandiam, B.; Corbu, A.; Massiot, G.; Sautel, F.; Yu, P.; Lin, B. W.-Y.; Houk, K. N.; Cossy, J. Intramolecular Diels–Alder Approaches to the Decalin Core of Verongidolide: The Origin of the exo-Selectivity, a DFT Analysis. *J. Org. Chem.* **2018**, *83*, 5975–5985.

(73) Byrne, M. J.; Lees, N. R.; Han, L.-C.; van der Kamp, M. W.; Mulholland, A. J.; Stach, J. E.; Willis, C. L.; Race, P. R. The catalytic mechanism of a natural Diels–Alderase revealed in molecular detail. *J. Am. Chem. Soc.* **2016**, *138*, 6095–6098.

(74) He, C. Q.; Chen, T. Q.; Patel, A.; Karabiyikoglu, S.; Merlic, C. A.; Houk, K. Distortion, tether, and entropy effects on transannular Diels–Alder cycloaddition reactions of 10–18-membered rings. *J. Org. Chem.* **2015**, *80*, 11039–11047.

(75) Yu, P.; Patel, A.; Houk, K. Transannular (6+ 4) and ambimodal cycloaddition in the biosynthesis of heronamide A. *J. Am. Chem. Soc.* **2015**, *137*, 13518–13523.

(76) Pham, H. V.; Paton, R. S.; Ross, A. G.; Danishefsky, S. J.; Houk, K. Intramolecular Diels–Alder reactions of cycloalkenones: stereoselectivity, Lewis acid acceleration, and halogen substituent effects. *J. Am. Chem. Soc.* **2014**, *136*, 2397–2403.

(77) Duan, A.; Yu, P.; Liu, F.; Qiu, H.; Gu, F. L.; Doyle, M. P.; Houk, K. Diazo esters as dienophiles in intramolecular (4+2) cycloadditions: Computational explorations of mechanism. *J. Am. Chem. Soc.* **2017**, *139*, 2766–2770.

(78) Fell, J. S.; Lopez, S. A.; Higginson, C. J.; Finn, M.; Houk, K. Theoretical Analysis of the Retro-Diels–Alder Reactivity of Oxanorbornadiene Thiol and Amine Adducts. *Org. Lett.* **2017**, *19*, 4504–4507.

(79) Levandowski, B. J.; Herath, D.; Gallup, N. M.; Houk, K. Origin of π-Facial Stereoselectivity in Thiophene 1-Oxide Cycloadditions. *J. Org. Chem.* **2018**, *83*, 2611–2616.

(80) Scholl, K.; Dillashaw, J.; Timpy, E.; Lam, Y.-h.; DeRatt, L.; Benton, T. R.; Powell, J. P.; Houk, K.; Morgan, J. B. Quinine-Promoted, Enantioselective Boron-Tethered Diels–Alder Reaction by Anomeric Control of Transition-State Conformation. *J. Org. Chem.* **2018**, *83*, 5756–5765.

(81) Suh, S.-E.; Chen, S.; Houk, K.; Chenoweth, D. M. The mechanism of the triple aryne–tetrazine reaction cascade: theory and experiment. *Chem. Sci.* **2018**, *9*, 7688–7693.

(82) Tan, D.; Jamieson, C. S.; Ohashi, M.; Tang, M.-C.; Houk, K.; Tang, Y. Genome-mined Diels–Alderase catalyzes formation of the cis-octahydrodecalins of varicidin A and B. *J. Am. Chem. Soc.* **2019**, *141*, 769–773.

(83) Schmidt, Y.; Lam, J. K.; Pham, H. V.; Houk, K.; Vanderwal, C. D. Studies on the Himbert intramolecular Arene/Allene Diels–alder cycloaddition. Mechanistic studies and expansion of scope to all-carbon tethers. *J. Am. Chem. Soc.* **2013**, *135*, 7339–7348.

(84) Liang, Y.; Mackey, J. L.; Lopez, S. A.; Liu, F.; Houk, K. Control and design of mutual orthogonality in biorthogonal cycloadditions. *J. Am. Chem. Soc.* **2012**, *134*, 17904–17907.

(85) Liu, S.; Lei, Y.; Qi, X.; Lan, Y. Reactivity for the Diels–Alder reaction of cumulenes: a distortion-interaction analysis along the reaction pathway. *J. Phys. Chem. A* **2014**, *118*, 2638–2645.

(86) Ma, Z.-X.; Patel, A.; Houk, K.; Hsung, R. P. Highly Torquoselective Electrocyclizations and Competing 1, 7-Hydrogen Shifts of 1-Azatrienes with Silyl Substitution at the Allylic Carbon. *Org. Lett.* **2015**, *17*, 2138–2141.

(87) Ohashi, M.; Liu, F.; Hai, Y.; Chen, M.; Tang, M.-c.; Yang, Z.; Sato,M.; Watanabe, K.; Houk, K.; Tang, Y. SAM-dependent enzyme-catalysed pericyclic reactions in natural product biosynthesis. *Nature* **2017**, *549*, 502–506.

(88) Yang, Z.; Dong, X.; Yu, Y.; Yu, P.; Li, Y.; Jamieson, C.; Houk, K. Relationships between product ratios in ambimodal pericyclic reactions and bond lengths in transition structures. *J. Am. Chem. Soc.* **2018**, *140*, 3061–3067.

(89) Xue, X.-S.; Jamieson, C. S.; Garcia-Borr`as, M.; Dong, X.; Yang, Z.; Houk, K. Ambimodal trispericyclic transition state and dynamic control of periselectivity. *J. Am. Chem. Soc.* **2019**, *141*, 1217–1221.

(90) Wiest, O.; Houk, K. Stabilization of the transition state of the chorismate-prephenate rearrangement: An ab initio study of enzyme and antibody catalysis. *J. Am. Chem. Soc.* **1995**, *117*, 11628–11639.

(91) Scott, S. K.; Sanders, J. N.; White, K. E.; Yu, R. A.; Houk, K.; Grenning, A. J. Controlling, understanding, and redirecting the thermal rearrangement of 3, 3-dicyano-1, 5-enynes. *J. Am. Chem. Soc.* **2018**, *140*, 16134–16139.

(92) Boon, B. A.; Green, A. G.; Liu, P.; Houk, K.; Merlic, C. A. Using ring strain to control 4π-electrocyclization reactions: torquoselectivity in ring closing of medium-ring dienes and ring opening of bicyclic cyclobutenes. *J. Org. Chem.* **2017**, *82*, 4613–4624.

(93) Patel, A.; Barcan, G. A.; Kwon, O.; Houk, K. Origins of 1, 6-Stereoinduction in Torquoselective 6π Electrocyclizations. *J. Am. Chem. Soc.* **2013**, *135*, 4878–4883.

(94) Sader, C. A.; Houk, K. A Theoretical Study of Cyclohexyne Addition to Carbonyl–Cα Bonds: Allowed and Forbidden Electrocyclic and Nonpericyclic Ring-Openings of Strained Cyclobutenes. *J. Org. Chem.* **2012**, *77*, 4939–4948.

(95) Hong, X.; Liang, Y.; Griffith, A. K.; Lambert, T. H.; Houk, K. Distortion-accelerated cycloadditions and strain-release-promoted cycloreversions in the organocatalytic carbonyl-olefin metathesis. *Chem. Sci.* **2014**, *5*, 471–475.

(96) Xie, S.; Lopez, S. A.; Ramstr¨om, O.; Yan, M.; Houk, K. 1, 3-Dipolar cycloaddition reactivities of perfluorinated aryl azides with enamines and strained dipolarophiles. *J. Am. Chem. Soc.* **2015**, *137*, 2958–2966.

(97) Krenske, E. H.; Patel, A.; Houk, K. Does nature click? Theoretical prediction of an enzyme-catalyzed transannular 1, 3-dipolar cycloaddition in the biosynthesis of lycojaponicumins A and B. *J. Am. Chem. Soc.* **2013**, *135*, 17638–17642.

(98) Celebi- Olcum, N.; Lam, Y.-h.; Richmond, E.; Ling, K. B.; Smith, A. D.; Houk, K. N. Pericyclic Cascade with Chirality Transfer: Reaction Pathway and Origin of Enantioselectivity of the Hetero-Claisen Approach to Oxindoles. *Angew. Chem. Intl. Ed.* **2011**, *50*, 11478–11482.

(99) Hamlin, T. A.; Kelly, C. B.; Ovian, J. M.; Wiles, R. J.; Tilley, L. J.; Leadbeater, N. E. Toward a unified mechanism for oxoammonium salt-mediated oxidation reactions: a theoretical and experimental study using a hydride transfer model. *J. Org. Chem.* **2015**, *80*, 8150–8167.

(100) Vitkovskaya, N. M.; Kobychev, V. B.; Bobkov, A. S.; Orel, V. B.; Schmidt, E. Y.; Trofimov, B. A. Nucleophilic Addition of Ketones To Acetylenes and Allenes: A Quantum-Chemical Insight. *J. Org. Chem.* **2017**, *82*, 12467–12476.

(101) Kister, J.; Ess, D. H.; Roush, W. R. Enantio-and Diastereoselective Synthesis of syn-β-Hydroxy-α-vinyl Carboxylic Esters via Reductive Aldol Reactions of Ethyl Allenecarboxylate with 10-TMS-9-Borabicyclo (3.3. 2) decane and DFT Analysis of the Hydroboration Pathway. *Org. Lett.* **2013**, *15*, 5436–5439.

(102) Johnson, E. R.; Clarkin, O. J.; Dale, S. G.; DiLabio, G. A. Kinetics of the addition of olefins to Si-centered radicals: the critical role of dispersion interactions revealed by theory and experiment. *J. Phys. Chem. A* **2015**, *119*, 5883–5888.

(103) Zhang, Z.; Collum, D. B. Wittig Rearrangements of Boron-Based Oxazolidinone Enolates. *J. Org. Chem.* **2019**, *84*, 10892–10900.

(104) Liu, W.-B.; Schuman, D. P.; Yang, Y.-F.; Toutov, A. A.; Liang, Y.; Klare, H. F.; Nesnas, N.; Oestreich, M.; Blackmond, D. G.; Virgil, S. C., et al. Potassium tert-butoxide-catalyzed dehydrogenative C–H silylation of heteroaromatics: a combined experimental and computational mechanistic study. *J. Am. Chem. Soc.* **2017**, *139*, 6867–6879.

(105) D'Alfonso, C.; Bietti, M.; DiLabio, G. A.; Lanzalunga, O.; Salamone, M. Reactions of the Phthalimide N-Oxyl Radical (PINO) with Activated Phenols: The Contribution of π-Stacking Interactions to Hydrogen Atom Transfer Rates. *J. Org. Chem.* **2013**, *78*, 1026–1037.

(106) DiLabio, G. A.; Franchi, P.; Lanzalunga, O.; Lapi, A.; Lucarini, F.; Lucarini, M.; Mazzonna, M.; Prasad, V. K.; Ticconi, B. Hydrogen atom transfer (HAT) processes promoted by the quinolinimide-N-oxyl radical. A kinetic and theoretical study. *J. Org. Chem.* **2017**, *82*, 6133–6141.

(107) Chan, B.; Gill, P. M.; Kimura, M. Assessment of DFT methods for transition metals with the TMC151 compilation of data sets and comparison with accuracies for main-group chemistry. *J. Chem. Theory Comput.* **2019**, *15*, 3610–3622.

(108) Chan, B. Assessment and development of DFT with the expanded CUAGAU-2 set of group-11 cluster systems. *Int. J. Quantum Chem.* **2021**, *121*, e26453.

(109) Frisch, M. J. et al. Gaussian 09 Revision A.1. Gaussian Inc. Wallingford CT 2009.

(110) Frisch, M. J. et al. Gaussian 16 Revision A.03. 2016; Gaussian Inc. Wallingford CT.

(111) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(112) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785.

(113) Yanai, T.; Tew, D. P.; Handy, N. C. A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(114) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(115) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(116) Johnson, E. R.; Becke, A. D. A post-Hartree-Fock model of intermolecular interactions: Inclusion of higher-order corrections. *J. Chem. Phys.* **2006**, *124*, 174104.

(117) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *J. Chem. Phys.* **1982**, *77*, 3654–3665.

(118) Hariharan, P. C.; Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor. Chim. Acta* **1973**, *28*, 213–222.

(119) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self-consistent molecular orbital methods. XII. Further extensions of Gaussian—type basis sets for use in molecular orbital studies of organic molecules. *J. Chem. Phys.* **1972**, *56*, 2257-2261.

(120) Mohamadi, F.; Richards, N. G.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. Macromodel—an integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comput. Chem.* **1990**, *11*, 440–467.

(121) Schrodinger Release 2020-3: MacroModel, Schr¨odinger, LLC. 2020; New York, NY.

(122) Schrodinger Release 2020-3: Maestro, Schr¨odinger, LLC. 2020; New York, NY.

(123) Medvedev, M. G.; Zeifman, A. A.; Novikov, F. N.; Bushmarinov, I. S.; Stroganov, O. V.; Titov, I. Y.; Chilov, G. G.; Svitanko, I. V. Quantifying possible routes for SpnF-catalyzed formal Diels–Alder cycloaddition. *J. Am. Chem. Soc.* **2017**, *139*, 3942–3945.

(124) Medvedev, M. G.; Panova, M. V.; Chilov, G. G.; Bushmarinov, I. S.; Novikov, F. N.; Stroganov, O. V.; Zeifman, A. A.; Svitanko, I. V. Exhaustive conformational search for transition states: the case of catechol O-methyltransferase active site. *Mendeleev Commun.* **2017**, *27*, 224–227.

(125) Fukaya, K.; Saito, A.; Nakajima, N.; Urabe, D. A computational study on the stereo-and regioselective formation of the C4α–C6′ bond of tethered catechin moieties by an exhaustive search of the transition States. *J. Org. Chem.* **2019**, *84*, 2840–2849.

(126) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(127) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys*. **1980**, *72*, 650–654.

(128) McLean, A.; Chandler, G. Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z= 11–18. *J. Chem. Phys.* **1980**, *72*, 5639–5648.

(129) Chang, G.; Guida, W. C.; Still, W. C. An internal-coordinate Monte Carlo method for searching conformational space. *J. Am. Chem. Soc.* **1989**, *111*, 4379–4386.

(130) https://github.com/bobbypaton/FullMonte.

(131) Kim, S.; Chmely, S. C.; Nimlos, M. R.; Bomble, Y. J.; Foust, T. D.; Paton, R. S.; Beckham, G. T. Computational study of bond dissociation enthalpies for a large range of native and modified lignins. *J. Phys. Chem. Lett.* **2011**, *2*, 2846–2852.

(132) Korth, M.; Pitonak, M.; Rezac, J.; Hobza, P. A transferable H-bonding correction for semiempirical quantum-chemical methods. *J. Chem. Theory Comput.* **2010**, *6*, 344–352.

(133) Stewart, J. J. P. MOPAC2016 (http://OpenMOPAC.Net). 2016; Colorado Springs, CO, USA.

(134) Neese, F. The ORCA program system. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 73–78.

(135) Neese, F. Software update: the ORCA program system, version 4.0. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8*, e1327.

(136) Dunning Jr, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(137) Kendall, R. A.; Dunning Jr, T. H.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

(138) Woon, D. E.; Dunning Jr, T. H. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *J. Chem. Phys.* **1993**, *98*, 1358–1371.

(139) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. Use of approximate integrals in ab initio theory. An application in MP2 energy calculations. *Chem. Phys. Lett.* **1993**, *208*, 359–363.

(140) Bernholdt, D. E.; Harrison, R. J. Large-scale correlated electronic structure calculations: the RI-MP2 method on parallel computers. *Chem. Phys. Lett.* **1996**, *250*, 477–484.

(141) Weigend, F.; H¨aser, M. RI-MP2: first derivatives and global consistency. *Theor. Chem. Acc.* **1997**, *97*, 331–340.

(142) Weigend, F.; K¨ohn, A.; H¨attig, C. Efficient use of the correlation consistent basis sets in resolution of the identity MP2 calculations. *J. Chem. Phys.* **2002**, *116*, 3175–3183.

(143) Vydrov, O. A.; Scuseria, G. E. Assessment of a long-range corrected hybrid functional. *J. Chem. Phys.* **2006**, *125*, 234109.

(144) Vydrov, O. A.; Heyd, J.; Krukau, A. V.; Scuseria, G. E. Importance of short-range versus long-range Hartree-Fock exchange for the performance of hybrid density functionals. *J. Chem. Phys.* **2006**, *125*, 074106.

(145) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. Design of density functionals by combining the method of constraint satisfaction with parametrization for thermochemistry, thermochemical kinetics, and noncovalent interactions. *J. Chem. Theory Comput.* **2006**, *2*, 364–382.

(146) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(147) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Constantin, L. A.; Sun, J. Workhorse semilocal density functional for condensed matter physics and quantum chemistry. *Phys. Rev. Lett.* **2009**, *103*, 026403.

(148) Chai, J.-D.; Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **2008**, *128*, 084106.

(149) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(150) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098.

(151) Perdew, J.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(152) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(153) Becke, A. D. A new mixing of Hartree–Fock and local density-functional theories. *J. Chem. Phys.* **1993**, *98*, 1372.

(154) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(155) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. Efficient, approximate and parallel Hartree–Fock and hybrid DFT calculations. A 'chain-of-spheres' algorithm for the Hartree–Fock exchange. *Chem. Phys.* **2009**, *356*, 98–109.

(156) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(157) Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.

(158) Goerigk, L.; Grimme, S. A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6670–6688.

(159) Becke, A. D.; Johnson, E. R. Exchange-Hole Dipole Moment and the Dispersion Interaction Revisited. *J. Chem. Phys.* **2007**, *127*, 154108.

(160) Otero-de-la-Roza, A.; Johnson, E. R. Non-Covalent Interactions and Thermochemistry using XDM-Corrected Hybrid and Range-Separated Hybrid Density Functionals. *J. Chem. Phys.* **2013**, *138*, 204109.

(161) The postg program is freely available from https://github.com/aoterodelaroza/postg.

(162) The canoical XDM coefficients can be found at http://schooner.chem.dal.ca and also in the xdm.param file of the postg distribution.

(163) Jensen, F. Polarization consistent basis sets: *Principles. J. Chem. Phys.* **2001**, *115*, 9113–9125.

(164) Jensen, F. Polarization consistent basis sets. II. Estimating the Kohn–Sham basis set limit. *J. Chem. Phys.* **2002**, *116*, 7372–7379.

(165) Jensen, F. Polarization consistent basis sets. III. The importance of diffuse functions. *J. Chem. Phys.* **2002**, *117*, 9234–9240.

(166) Jensen, F.; Helgaker, T. Polarization consistent basis sets. V. The elements Si–Cl. *J. Chem. Phys.* **2004**, *121*, 3463–3470.

(167) Karton, A. Highly accurate CCSDT(Q)/CBS reaction barrier heights for a diverse set of transition structures: basis set convergence and cost-effective approaches for estimating post-CCSD(T) contributions. *J. Phys. Chem. A* **2019**, *123*, 6720–6732.

(168) East, A. L.; Allen, W. D. The heat of formation of NCO. *J. Chem. Phys.* **1993**, *99*, 4638–4650.

(169) East, A. L.; Johnson, C. S.; Allen, W. D. Characterization of the ˜X 1A′ state of isocyanic acid. *J. Chem. Phys.* **1993**, *98*, 1299–1328.

(170) Papajak, E.; Truhlar, D. G. What are the most efficient basis set strategies for correlated wave function calculations of reaction energies and barrier heights? *J. Chem. Phys.* **2012**, *137*, 064110.

(171) Warden, C. E.; Smith, D. G.; Burns, L. A.; Bozkaya, U.; Sherrill, C. D. Efficient and automated computation of accurate molecular geometries using focal-point approximations to large-basis coupled-cluster theory. *J. Chem. Phys.* **2020**, *152*, 124109.

(172) Neese, F.; Valeev, E. F. Revisiting the atomic natural orbital approach for basis sets: Robust systematic basis sets for explicitly correlated and conventional correlated ab initio methods? *J. Chem. Theory Comput*. **2011**, *7*, 33–43.

(173) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. Basis-set convergence of correlated calculations on water. *J. Chem. Phys. 1997*, **106**, 9639–9646.

(174) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. Basis-set convergence in correlated calculations on Ne, N2, and H2O. *Chem. Phys. Lett.* **1998**, *286*, 243–252.

(175) Truhlar, D. G. Basis-set extrapolation. *Chem. Phys. Lett.* **1998**, *294*, 45–48.

(176) Burns, L. A.; Marshall, M. S.; Sherrill, C. D. Appointing silver and bronze standards for noncovalent interactions: A comparison of spin-component-scaled (SCS), explicitly correlated (F12), and specialized wavefunction approaches. *J. Chem. Phys.* **2014**, *141*, 234111.

(177) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. Benchmark database of accurate (MP2 and CCSD (T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(178) Rezac, J.; Hobza, P. Describing noncovalent interactions beyond the common approximations: How accurate is the "gold standard," CCSD (T) at the complete basis set limit? *J. Chem. Theory Comput.* **2013**, *9*, 2151–2155.

(179) Karton, A.; Martin, J. M. Comment on: Estimating the Hartree–Fock limit from finite basis set calculations (Jensen F (2005) Theor Chem Acc 113: 267). *Theor. Chem. Acc.* **2006**, *115*, 330–333.

(180) Jensen, F. Estimating the Hartree—Fock limit from finite basis set calculations. *Theor. Chem. Acc.* **2005**, *113*, 267–273.

(181) Mardirossian, N.; Head-Gordon, M. How accurate are the Minnesota density functionals for noncovalent interactions, isomerization energies, thermochemistry, and barrier heights involving molecules composed of main-group elements? *J. Chem. Theory Comput.* **2016**, *12*, 4303–4325.

(182) Goerigk, L.; Kruse, H.; Grimme, S. Benchmarking density functional methods against the S66 and S66x8 datasets for non-covalent interactions. *Chem. Phys. Chem* **2011**, *12*, 3421–3433.

(183) Grimme, S.; Huenerbein, R.; Ehrlich, S. On the importance of the dispersion energy for the thermodynamic stability of molecules. *Chem. Phys. Chem* **2011**, *12*, 1258–1261.

(184) Grimme, S. Seemingly simple stereoelectronic effects in alkane isomers and the implications for Kohn–Sham density functional theory. *Angew. Chem. Intl. Ed.* **2006**, *45*, 4460–4464.

(185) Goerigk, L. How do DFT-DCP, DFT-NL, and DFT-D3 compare for the description of London-dispersion effects in conformers and general thermochemistry? *J. Chem. Theory Comput.* **2014**, *10*, 968–980.

(186) Holmes, J. D.; Otero-de-la-Roza, A.; DiLabio, G. A. Accurate modeling of water clusters with density-functional theory using atom-centered potentials. *J. Chem. Theory Comput.* **2017**, *13*, 4205–4215.

Part IV

In this part, an investigation was carried out to examine the extent to which ACPs can overcome the deficiencies of minimal or double-ζ basis set Hartree−Fock (HF) methods. For this purpose, a more extensive and diverse set of non-covalent properties (105,880 entries) was utilized compared to the earlier proof-of-concept study. The target elements were also increased from four (H, C, N, O) to ten (H, B, C, N, O, F, Si, P, S, Cl). The diverse pool of model systems, including those from the newly generated PEPCONF data set, served as a good test for ACPs for various chemical environments that were not tested in the earlier work and helped identify the strengths and weaknesses associated with the ACP-based approaches.

Another goal of the presented work was to compare the performance of the ACP-based approaches with two other semi-empirical correction schemes from the literature. These two semi-empirical correction schemes were designed to correct the missing dispersion from HF methods and the basis set incompleteness error. Therefore, an examination of the effects of applying ACPs and the two semi-empirical correction schemes to minimal or double-ζ basis set HF methods provided an understanding of their relative merits. The objective behind the comparative study was also to identify if there was any value in developing ACPs for use with one or more of the semi-empirical correction schemes so that the underlying errors could be reduced further.

The primary data that support the findings of Chapter 7 is provided in Appendix 4 of this dissertation. Other supporting files have also been deposited to the figshare repository and are openly available at the following URL/DOI: https://doi.org/10.6084/m9.figshare.16912201. The reference of the published paper is as follows: Prasad, V. K.; Otero-de-la-Roza, A.; DiLabio, G. A. *Elec. Struc.* 2021, 3 (3), 034007. © Copyright 2021 Institute of Physics Publishing Ltd. (DOI: 10.1088/2516-1075/ac22b8)

Chapter 7

Performance of small basis set Hartree–Fock methods for modeling non-covalent interactions

Abstract

Non-covalent interactions (NCIs) play an essential role in (bio)chemistry. Wavefunction-based methods combined with large basis sets are able to accurately describe inter-and intra-molecular NCIs but are not practical for large molecular systems. Semi-empirical corrections have been developed recently that, when combined with Hartree–Fock (HF) and a small basis set, show promise in the ability to predict non-covalent binding and conformational energies over a wide range of systems. Compared to large-basis-set correlated wavefunction methods, small-basis-set HF methods have significantly lower computational cost and are useful for modeling large molecular systems with sizes between many hundred and a few thousand atoms. Using a large collection of non-covalent binding energies, conformational energies, and molecular deformation energies containing 105,880 entries, we provide a comprehensive evaluation of the performance of minimal basis set (MINIX) HF method with three correction schemes: D3, 3c, and atom-centered potentials (ACPs). We also evaluate the performance of HF/6-31G* in combination with the D3 and ACP schemes. By comparing the three corrections, we analyze the strengths and weaknesses associated with each strategy in predicting NCIs. Our results show that D3 corrections alone do not offer significant improvements in the performance of HF/MINIX or HF/6-31G* and, in some cases, overestimate binding energies resulting in large errors when compared to the reference data. The correction strategies that offer the best reduction in the underlying errors of HF/MINIX and HF/6-31G* are shown to be 3c and ACP for HF/MINIX and ACP for HF/6-31G*.

1. Introduction

Non-covalent interactions[1–3] (NCIs) are an important topic in various areas of physics, chemistry, biology, and materials science. NCIs occur in self-assembly, molecular recognition, supramolecular host-guest binding, crystal packing, and many other contexts.[4] They play a crucial role not just at the microscopic level, e.g., in determining the structure, dynamics, and function of biomolecules, but also at a macroscopic level, e.g., the formation of liquid and solid phases of matter.[5–8] NCIs, which include electrostatic, induction, dispersion, etc. contributions, vary in strength and operate over different length scales.[9–11]

Quantum mechanical (QM) methods are an important tool for modeling NCIs.[12–17] However, an accurate QM description of NCIs typically requires large basis sets and a high-level treatment of electron correlation, the latter correcting for the absence of dispersion in the underlying Hartree–Fock (HF) method.[18] Such calculations are expensive and only applicable to relatively small molecular systems.[19] Consequently, the development of new QM methods that can efficiently and accurately predict NCIs is an active field of research.

In 2010, Grimme and co-workers proposed their popular D3 dispersion correction for Hartree–Fock (HF) and density functional theory (DFT) methods.[20–22] D3 is a semi-empirical energy correction to account for the missing dispersion energy in HF and DFT methods. However, because basis set incompleteness effects impact negatively the calculation of NCIs[23–26], D3-corrected HF or DFT methods require large basis sets to be accurate. As such calculations scale approximately as the third power of the number of basis functions, they are impractical for large systems.

The high computational cost associated with modeling large molecular systems using large-basis-set post-HF or DFT methods has driven the development of low-cost QM alternatives.[27–30] Some of these methods take advantage of the efficiency of the HF method with small basis sets while introducing semi-empirical corrections parametrized with experimental data to mitigate the performance shortcomings caused by absence of correlation and basis set incompleteness.[31–34] In 2013, Sure *et al.* proposed a nine-parameter semi-empirical correction called 3c.[35] The HF-3c method uses a small basis set that is minimal for first row atoms (MINIX). In HF-3c, the basis set incompleteness error is mitigated with a geometrical counterpoise correction (gCP), and dispersion is treated using D3 (with parameters different from complete-basis-set HF). A third term, called the short-range basis (SRB) incompleteness correction, was introduced to fix the spuriously long covalent bond lengths when employing a small basis set, particularly with electronegative elements (N, O, F, etc.).

A different approach adopted in our research group is based on atom-centered potentials (ACPs).[36–45] ACPs are one-electron potentials with the same mathematical formulation as effective-core potentials[47,48] and allow for an economical means of correcting the underlying shortcomings of QM methods. Earlier generations of ACPs were specifically designed to incorporate the dispersion interactions missing in conventional DFT methods.[36–43] More recently, we demonstrated that ACPs could also be used to correct basis set incompleteness in DFT (ACPs developed for this purpose are called basis set incompleteness potentials, BSIPs).[24,25] We also showed that ACPs developed for, and used with, minimal-basis-set HF are capable of reproducing non-covalent binding and conformational energies obtained with complete-basis-set CCSD(T) accurately and at a small fraction of the computational cost.[39]

In this work, we show how the different corrections can be applied to small-basis-set HF to model NCIs in large molecular systems and analyze their performance and relative merits. We first explore the performance of the HF method in conjunction with a minimal basis set, i.e., MINIX, and a double-ζ basis set, i.e., 6-31G*. We then apply to HF/MINIX three different correction strategies and examine their effectiveness: D3, 3c, and ACP. We also look at the effects of the application of D3 and ACP on HF/6-31G* (no 3c parameters are available for HF/6-31G*). The performance of D3, 3c, and ACP corrected HF regarding NCIs is examined using various benchmark datasets from the literature and the strengths and weaknesses associated with each strategy are identified.

2. Methodology

2.1 Theoretical background

The first correction strategy considered in this work is the D3 dispersion-correction scheme[20–22] (with Becke-Johnson or BJ damping[49]). This correction is calculated as:

$$E^{D3(BJ)} = -\frac{1}{2}\sum_{A\neq B}^{atoms}\left(s_6\frac{C_6^{AB}}{R_{AB}^6+(a_1R_{AB}^0+a_2)^6}+s_8\frac{C_8^{AB}}{R_{AB}^8+(a_1R_{AB}^0+a_2)^8}\right) \qquad (1)$$

where C_6^{AB} and C_8^{AB} are the sixth-and eight-order dispersion coefficients for atom pairs A and B, s_6 and s_8 are scaling factors, a_1 and a_2 are fitted parameters, R_{AB} is the interatomic distance between atoms A and B, and R_{AB}^0 is equal to $(C_8^{AB}/C_6^{AB})^{1/2}$. The datasets used in the fitting procedure to obtain the HF parameters in Equation 1 were: S22, S22+, PCONF, SCONF, ACONF, CCONF, ADIM6, and RG6 (see reference 22 and references therein). The D3 correction scheme depends only on the atomic coordinates and the four pre-determined parameters, s_6, s_8, a_1, and a_2. Therefore, the computational cost associated with determining $E^{D3(BJ)}$ is negligible in comparison with the underlying HF calculation. For HF, the values of the parameters in Equation 1 are: $s_6 = 1.0$, $s_8 = 0.9171$, $a_1 =$ 0.3385, and $a_2 = 2.8830$ Å.

The second correction, the HF-3c approach, combines the aforementioned D3 term (with refitted parameters) with two additional terms, gCP and SRB. The refitted parameters for D3 in the HF-3c approach were obtained using the high-level non-covalent binding energies of the S66 dataset, yielding: $s_6 = 1.0$, $s_8 = 0.8777$, $a_1 =$ 0.4171, and $a_2 = 2.9149$ Å.[35] The gCP[50] term is a geometrical counterpoise correction designed to mitigate basis set superposition error. It is calculated as:

$$E^{gCP} = \sigma \sum_{A}^{atoms} \sum_{A \neq B}^{atoms} E_A^{miss} \frac{\exp(-\alpha(R_{AB})^{\beta})}{\sqrt{S_{AB} N_B^{virt}}} \tag{2}$$

where σ is a global scaling factor, E_A^{miss} refers to the pre-computed atomic energy difference between a nearly complete basis set and the target basis set (MINIX in this case[35]), $\exp(-\alpha(R_{AB})^{\beta})$ acts as a decay function that depends on the interatomic distance R_{AB} between atom pairs (A and B) and the fitted parameters α and β, and the $(S_{AB} N_B^{virt})^{-1/2}$ term is a normalization constant that depends on Slater-type overlap integrals S_{AB} and the number of virtual orbitals N_B^{virt} on atom B. The S_{AB} integrals are further dependent on a fitted parameter η. The parameters for gCP were obtained by fitting to the non-covalent binding energies in the S66x8 dataset.[35] The gCP energy correction depends on four parameters (σ, α, β, η) and can be calculated from the geometry alone, so it has a negligible computational cost.

The third component of HF-3c is SRB[35], a short-range basis incompleteness correction intended to correct for the spuriously long covalent bond lengths predicted by HF/MINIX. SRB has the following form:

$$E^{SRB} = -s \sum_{A}^{atoms} \sum_{A \neq B}^{atoms} (Z_A Z_B)^{3/2} \exp(-\gamma \left(R_{AB}^{0,D3}\right)^{3/4} R_{AB}) \tag{3}$$

where Z_A and Z_B are nuclear charges associated with the atoms A and B, s and γ are fitted parameters with values 0.03 and 0.7, R_{AB} is the interatomic distance, and $R_{AB}^{0,D3}$ are the cutoff radii for the D3 dispersion correction scheme. The values of the fitted parameters s and γ were determined by fitting against high-level atomic forces in a set of 107 equilibrium structures of small organic molecules. The SRB energy can also be calculated from the geometry alone and has negligible computational cost.

When small-basis-set HF is corrected with one of the methods above, the total energy is:

$$E_{total}^{HF-D3} = E^{HF} + E^{D3(BJ)} \tag{4}$$

$$E_{total}^{HF-3c} = E^{HF} + E^{D3(BJ)} + E^{gCP} + E^{SRB} \tag{5}$$

The other correction strategy examined in this article involves the use of ACPs. The ACP development procedure was described in detail in previous works[25,39]. ACPs are similar to effective-core potentials[47,48] except they do not replace any electrons. They are represented in potential operator form as:

$$\hat{V}_{ACP}(r) = \sum_{\alpha}\left(V_{local}^{\alpha}(r) + \sum_{l=0}^{L-1}\sum_{\mathrm{m}=-l}^{l}\delta V_{l}^{\alpha}(r)\,|Y_{lm}\rangle\langle Y_{lm}|\right) \tag{6}$$

where $\delta V_l^{\alpha}(r) = V_l^{\alpha}(r) - V_{local}^{\alpha}(r)$, α represents atoms on which potentials are centered, and r is the distance to atom α. $|Y_{lm}\rangle\langle Y_{lm}|$ are projection operators using real spherical harmonics based on atom α with angular momentum quantum number l and magnetic quantum number m. The $V_{local}^{\alpha}(r)$ and $\delta V_l^{\alpha}(r)$ terms in Equation 6 are written using Gaussian functions:

$$V_l^{\alpha}(r) = \sum_{n=1}^{N} c_{ln}^{\alpha}\exp(-\xi_{ln}^{\alpha}r^2) \qquad for\;\; l = 0, 1, 2, \dots, L \tag{7}$$

where N is the total number of Gaussian-type functions. The coefficients (c_{ln}^{α}) and exponents (ξ_{ln}^{α}) are adjustable parameters determined via regularized least-squares fitting to reference data.

To correct the underlying errors of HF/MINIX or HF/6-31G* using ACPs (the corrected methods are referred to as HF/MINIX-ACP or HF/6-31G*-ACP), the ACP operator from Equation 6 is added as a one-electron potential to HF. To first order in the ACP perturbation, the energy is:

$$E_{total}^{HF-ACP} = E^{HF} + E^{ACP}(c, \xi) \tag{8}$$

The first-order ACP energy correction is:

$$\begin{aligned} E^{ACP}(c,\xi) &= \sum_{i}\langle\psi_i|\hat{V}_{ACP}(r)|\psi_i\rangle \\ &= \sum_{\alpha ln} c_{ln}^{\alpha}\sum_{i}\langle\psi_i|(\,|Y_{lm}\rangle\exp(-\xi_{ln}^{\alpha}r^2)\,\langle Y_{lm}|\,)|\psi_i\rangle \\ &= \sum_{\alpha ln} c_{ln}^{\alpha}\,\Delta E_{ln}^{\alpha}(\xi_{ln}^{\alpha}) = \boldsymbol{c}\cdot\boldsymbol{\Delta}\mathbf{E}(\boldsymbol{\xi})^{T} \end{aligned} \tag{9}$$

where index i runs over occupied orbitals ψ. The $\boldsymbol{c}$ and $\boldsymbol{\Delta}\mathbf{E}(\boldsymbol{\xi})^{T}$ are ACP coefficient and ACP energy term vectors, respectively.

The ACP development process involves compiling a large and diverse training set comprising the target molecular properties. The molecules in the training set contain the atoms for which the ACPs are being developed. For each entry in the training set, the ACP terms ($\Delta E_{ln}^{\alpha}(\xi_{ln}^{\alpha})$ from Equation 9) are computed for a pre-determined set of exponents and angular momenta on each atom. Once the effect of each ACP term has been calculated, the optimal ACP coefficients and associated exponents are determined using a regularized least-squares fitting procedure subject to a constraint on the 1-norm (sum of the absolute values) of the ACP coefficients. The ACP development process culminates with the generation of ACPs that correct the underlying errors in systems that closely resemble the training set. Compared to the underlying HF method, ACPs add approximately 10–30% to the overall calculation time in the *Gaussian16*[51] program. ACPs must also be developed for each atom for which their use is intended.

2.2 Test sets

To compare the performance of HF/MINIX or HF/6-31G* (with and without D3 or ACP corrections) and HF-3c (which uses the MINIX basis set), we have used two classes of test sets: The "training set" is a collection of datasets used to parameterize the ACPs, and the "validation set" is a collection of datasets used to test the accuracy of properties computed using ACPs for systems not included in the training set. It should be kept in mind that performance comparisons on the training set will be biased in favor of the ACPs, but comparisons based on the validation set will be more balanced. As the focus of this work is NCIs, the training and validation sets include various datasets representing non-covalent binding energies, conformational energies, and deformation energies relative to the molecular equilibrium geometry. Note that any data point involving molecules containing atoms other than H, B, C, N, O, F, Si, P, S, and Cl – that is, the atoms for which ACPs were developed – were not included in the training or validation sets. The list of datasets in the training and validation sets is given in Tables 1 and 2. The datasets have been also categorized to facilitate the presentation of the results. In total, the training set comprises 19,439 non-covalent binding energies, 44,105 conformational energies, and 10,288 molecular deformation energies. The validation set contains 27,811 binding energies and 4,237 conformational energies. The majority of the reference data used in both training and validation sets was calculated with complete-basis-set wavefunction theory methods, with exceptions noted in Tables 1 and 2.

Table 1. List of datasets in the training set. (The "*Subset*" column refers to the various grouping of data sets done based on the features of the data points. The "*Dataset(s)*", "*Datapoints*", "*Description*", and "*References*" columns respectively indicate the names of the data sets that comprise the defined subset, the total number of data points in the subset, a brief description of the feature of data points, and the references of the data sets comprising the subset.)

Subset	Data set(s)	Data points	Description	References
Non-covalent binding energies of molecular complexes:				
Mixed NCIs	S22x5, S66x8, S66a8, A21x12, NBC10ext, Sulfurx8, 3B-69-DIM, 3B-69-TRIM, WatAA[a]	2,228	Mix character non-covalent interactions	63,64,73–75,65–72

Subset	Data set(s)	Data points	Description	References
Protein-Protein	BBI, SSI	2,905	Interactions between proteins	76
Nucleotide-Nucleotide	JSCH, DNAstack, DNA2body, ACHC, BDNA, NucBTrimer[a]	440	Interactions between nucleotides	63,77–80
Nucleotide-Protein	NucTAA[a]	454	Interactions between nucleotide and protein	81–84
Carbohydrate-Protein	CarbhydBz, CarbhydNaph, CarbhydAroAA[a], CarbhydAro[a]	289	Interactions between carbohydrate and protein	85–88
Biomolecule-Drug	HSG, PLF547	409	Interactions between drugs and biomolecules	71,89,90
Hydrogen-bonding	HBC6, MiriyalaHB104, IonicHB, HB375x10, IHB100x10, HB300SPXx10	6,397	Hydrogen bonding interactions	71,91–96
π-stacking	Pisub[a], Pi29n, BzDC215	304	Non-stacked and stacked π-π interactions	97–100
Halogen-bonding	Hill18, X40x10	238	Halogen bonding interactions	101,102
Other-NCI	PNICO23, CARBH12	35	Interactions in pnicogen-bonded systems and singlet carbene systems	103,104
Hydrocarbon-BE	ADIM6, HC12, HW30, C2H4NT	123	Aliphatic-aliphatic interactions	22,69,105,106
Gas-adsorption	CH4PAH, CO2MOF, CO2PAH, CO2NPHAC, BzGas	876	Interactions between gas and substrate molecules	69,107–111
Water-BE1	Water38, Water1888, Water-2body[e]	2,336	Hydrogen-bonded water dimers and $(H_2O)_n$ clusters where n=3-10	46,69,112–115
BFSiPSCl	B-set[a], F-set[a], Si-set[a], P-set[a], S-set[a], Cl-set[a]	896	Monomers contain B, F, Si, P, S, or Cl atoms	25
Anionic-BE1[b]	SSI, WatAA[a], HSG, PLF547, IonicHB, IHB100x10, Ionic43	1,509	Anionic interactions	75,76,89,90,94,95,116
Molecular conformational energies:				
Protein	PEPCONF-Dipeptide[a], TPCONF, P76, YMPJ	1,450	Peptide-like model systems	117–120
DNA	SPS	17	DNA-like model systems	121
RNA	rSPS, UpU46	90	RNA-like model systems	122,123
Carbohydrate	SCONF, DSCONF, SacchCONF, CCONF	526	Carbohydrate-like model systems	104,124–127
Hydrocarbon-REL	ACONF, BCONF, PentCONF	421	Hydrocarbon-like model systems	128–130
Water-REL	Undecamer125	124	$(H_2O)_{11}$ clusters	131
Miscellaneous	ICONF, MCONF, Torsion21, 37Conf8, DCONF, MolCONF	8,280	Various small molecules	104,132–136
ANI1ccxCONF	ANI1ccxCONF[c]	32,944	Various organic molecules	137
Anionic-REL1[d]	PEPCONF-Dipeptide[a], MolCONF	254	Negatively charged molecules	117,136
Molecular deformation energies:				
MOLdef	MOLdef[a]	9,298	Various small molecules deformed along their normal modes	25

Subset	Data set(s)	Data points	Description	References
MOLdef-H2O	MOLdef-H2O[e]	990	Various H_2O systems deformed along their normal modes	138,139

[a)] the reference data was calculated as part of a different work at DLPNO-CCSD(T)/CBS level, on geometries from the literature using a basis set extrapolation scheme described in detail in reference 53

[b)] comprises non-covalently bound complexes, where at least one of the monomers is negatively charged (it contains systems from other datasets except for NucTAA)

[c)] contains majorly conformational energies but also some molecular deformation energies

[d)] comprises negatively charged conformers (it contains systems from other datasets except for SPS, rSPS, and UpU46)

[e)] the reference data has been calculated as part of a different work (see reference 53) at CCSD(T)/CBS level using the same extrapolation scheme as described in reference 112

Table 2. List of datasets in the validation set. (The "*Subset*" column refers to the various grouping of data sets done based on the features of the data points. The "*Dataset(s)*", "*Datapoints*", "*Description*", and "*References*" columns respectively indicate the names of the data sets that comprise the defined subset, the total number of data points in the subset, a brief description of the feature of data points, and the references of the data sets comprising the subset.)

Subset	Dataset(s)	Datapoints	Description	References
Non-covalent binding energies of molecular complexes:				
BlindNCI	BlindNCI	80	Mix character non-covalent interactions	140
DES15K	DES15K	11,474	Mix character non-covalent interactions	141
NENCI-2021	NENCI-2021	5,859	Mix character non-covalent interactions	142
R160X6	R160x6	960	Repulsive interactions	143
R739X5	R739x5	4,330	Repulsive interactions	144
CE20	CE20	20	Hydrogen bonding interactions	145,146
CHAL336	CHAL336	48	Chalcogen bonding interactions	147
XB45	XB45	33	Halogen bonding interactions	148
WaterOrg	WaterOrg	2,376	Hydrogen bonding interactions between water clusters and organic molecules	149
Water-BE2[a]	Water27, HW6Cl, HW6F, FmH2O10, SW49Bind345, SW49Bind6, H2O20Bind10	107	Hydrogen bonding interactions in various water cluster systems	150–154
L7	L7	7	Interactions in large molecules	155,156
S12L	S12L	10	Interactions in large molecules	156–158
S30L	S30L	26	Interactions in large molecules	159
C60dimer	C60dimer	14	Interactions in C_{60} dimers	160

Subset	Dataset(s)	Datapoints	Description	References
Ni2021	Ni2021	11	Interactions in large molecules	161
Anionic-BE2[b]	Anionpi, IL236, DES15K, NENCI-2021, CHAL336, XB45, S30L	2,455	Anionic interactions	141,142,147,148,159,162,163
Molecular conformational energies:				
SafroleCONF	SafroleCONF	5	Safrole or 5-(2-propenyl)-1,3-benzodioxol)	164
AlcoholCONF	AlcoholCONF	31	Small alcohols	165
BeranCONF	BeranCONF	50	Small organic molecules	166
Torsion30	Torsion30	2,107	Biaryl drug-like molecules	167
MPCONF196	MPCONF196	112	Macrocyclic peptide model systems	168
PEPCONF-Tripeptide	PEPCONF-Tripeptide[c]	647	Tripeptide model systems	117
PEPCONF-Disulfide	PEPCONF-Disulfide[d]	620	Disulfide-bridged peptide model systems	117
PEPCONF-Cyclic	PEPCONF-Cyclic[d]	320	Macrocyclic peptide model systems	117
PEPCONF-Bioactive	PEPCONF-Bioactive[d]	175	Peptide model systems with associated bioactivity	117
Anionic-REL2	PEPCONF-Disulfide[d], PEPCONF-Bioactive[d]	170	Negatively charged molecules	117

[a] this category comprises datasets with water clusters, neutral and charged.

[b] comprises non-covalently bound complexes, with at least one monomer negatively charged.

[c] only the subset from the PEPCONF[117] database for which reference data was recalculated at DLPNO-CCSD(T)/CBS level (see reference 53).

[d] available reference data was calculated at LC-ωPBE-XDM/aug-cc-pVTZ level of theory.

2.3 Technical details

All the self-consistent field (SCF) HF/MINIX and HF/6-31G* calculations, including those that are D3- and ACP-corrected, were performed with the *Gaussian16*[51] software package. The HF-3c energies were obtained from the *ORCA*[52] software package. The usage of ACPs with HF/MINIX or HF/6-31G* in *Gaussian16* is demonstrated in the Supporting Information (SI) via a sample input file. Additional detailed information related to the development of ACPs can be found in reference 53.

We developed ACPs for ten atoms (H, B, C, N, O, F, Si, P, S, and Cl). ACP terms were calculated for angular momentum channels up to the maximum angular momentum of the corresponding basis set for the corrected atoms (α): *s* for H (MINIX, 6-31G*), *p* for B, C, N, O, F (MINIX), *d* for Si, P, S, Cl (MINIX), and *d* for B, C, N, O, F, Si, P, S, Cl (6-31G*). We pre-selected 29 ACP exponents (ξ_{ln}^{α}, see Equation 7): 0.12 to 0.30 in 0.02 steps, 0.40 to 2.00 in 0.10 steps, and 2.50 to 3.00 in 0.50 steps. The total number of

ACP terms were 957 for MINIX and 1102 for 6-31G*. The ACP term generation and the ACP fitting were carried out using the *dcp*[54] scripts and *acpfit*[55] program, which are publicly available[56]. To obtain the optimized ACP coefficient and exponent pairs (see Equation 7), we used the Least Absolute Shrinkage and Selection Operator[57] (LASSO) regression technique for performing the constrained least-squares fitting with constraint on the 1-norm of the coefficients. The optimized ACPs generated in this work for HF/MINIX and HF/6-31G* are provided in the SI.

3. Results and Discussion

The results of evaluating the HF/MINIX and HF/6-31G* (with and without D3 or ACP corrections) and HF-3c methods with the training and validation sets are shown in Figures 1 to 4. A detailed breakdown of the errors associated with each method by subset can be found in the SI. Note that the D3 and 3c corrections were parametrized using a small subset of the datasets presented here (see Section 2.1).[22,35,50] Therefore, the results presented in this work for both the training and validation sets are good tests of both HF-D3 and HF-3c.

3.1 Performance of HF/MINIX-based methods on the training set

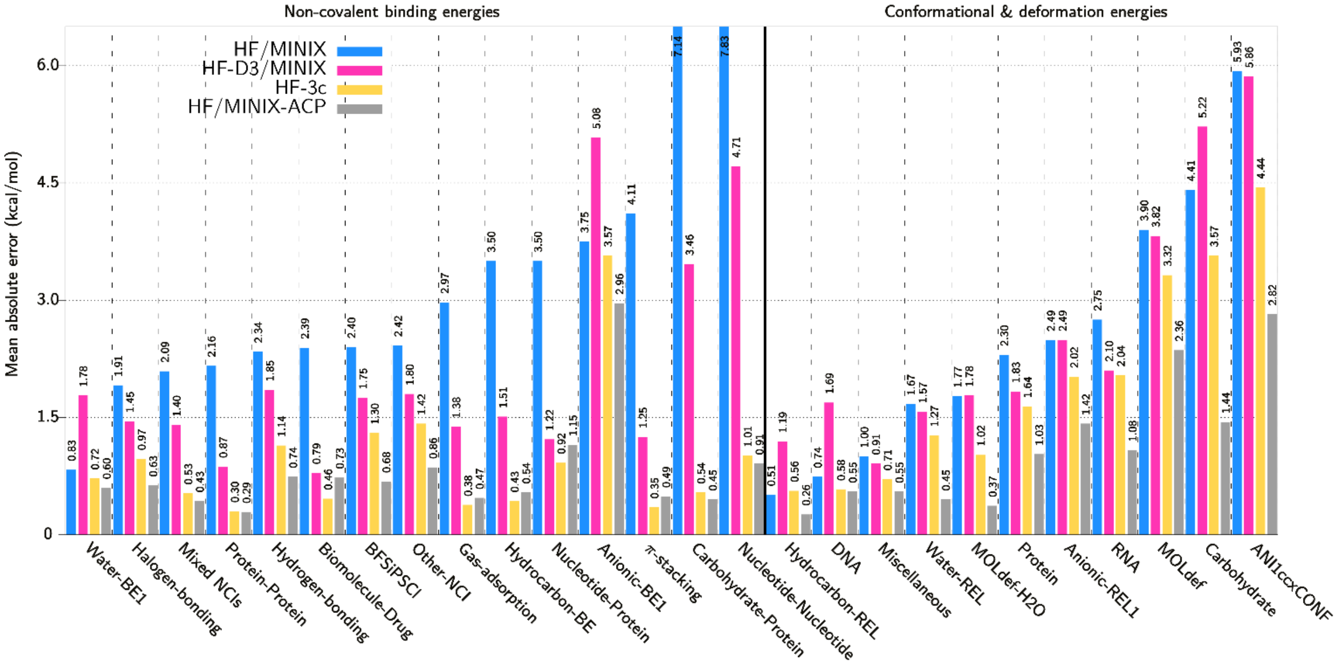


Figure 1. Mean absolute errors of HF/MINIX-based methods (relative to the reference data) for the training set (Table 1). The methods shown are HF/MINIX (blue), HF-D3/MINIX (pink), HF-3c (yellow), and HF/MINIX-ACP (grey). The values for the mean absolute errors (in kcal/mol) are given atop the bars.

The mean absolute errors (MAEs) of HF/MINIX for the training set properties are presented in Figure 1. For the non-covalent binding energy subsets (left of the vertical bar in Figure 1), the MAEs range from 0.83 to 7.83 kcal/mol. The lowest MAE is obtained for the subset containing small water clusters (*Water-BE1* subset). Other subsets representing NCIs with dominant electrostatic character, such as *Halogen-bonding, Hydrogen-bonding,* and *Other-NCI,* also have large but reasonable errors, with MAEs between 1.91 to 2.42 kcal/mol. However, NCI types that are either dominated by dispersion, such as aliphatic-aliphatic interactions (*Hydrocarbon-BE*, MAE=3.50 kcal/mol) and π-π interactions (*π-stacking*, MAE=4.11 kcal/mol) or by anion-containing species (*Anionic-BE1*, MAE=3.75 kcal/mol) have much higher errors, comparable to the absolute value of the binding energies themselves. The large errors for individual NCI types render HF/MINIX unusable for applications involving large molecules. This is demonstrated by its poor performance in subsets representing NCIs of biological relevance like *Mixed NCIs*, *Protein-Protein*, *Nucleotide-Protein, Carbohydrate-Protein, Nucleotide-Nucleotide,* and *Biomolecule-Drug,* with MAEs between 2.09 to 7.83 kcal/mol. Errors for gas adsorption model systems (*Gas-adsorption* subset), another potentially interesting context in which these low-cost methods can be used, are also quite large (MAE=2.97 kcal/mol). For the conformational and deformation energy subsets (right of the vertical bar in Figure 1), which are used to gauge the performance of methods for screening of conformers and structural predictions, the MAEs of HF/MINIX vary between 0.51 to 5.93 kcal/mol. A very low MAE is obtained for hydrocarbon conformers (*Hydrocarbon-REL* subset) indicating that HF/MINIX without any applied correction is surprisingly adequate for conformational searching of hydrocarbon-like systems. As seen from Figure 1, the performance of HF/MINIX is worst for the non-covalent binding energies in subsets *Nucleotide-Nucleotide* and *Carbohydrate-Protein,* where it predicts the complexes to be under-bound or repulsive, resulting in MAEs of 7.83 and 7.14 kcal/mol, respectively. One common feature of the systems in the *Nucleotide-Nucleotide* and *Carbohydrate-Protein* subsets is that all complexes contain at least one delocalized aromatic monomer, resulting in some form of stacked π-π interactions, non-stacked π-π interactions, or H-π interactions. The failure of HF/MINIX to properly model aromatic complexes is further underscored by the MAE of 4.11 kcal/mol for the subset of π-π interactions (*π-stacking*). HF/MINIX also lacks the ability to accurately predict non-covalent binding and conformational energies of subsets like *Anionic-BE1* (MAE=3.75 kcal/mol) and *Anionic-REL1* (MAE=2.49 kcal/mol), whose systems contain at least one negatively charged molecule. These systems are challenging for HF/MINIX because MINIX lacks the diffuse basis functions required for the proper description of negatively charged molecule(s). Among all conformational energy subsets of biological relevance, HF/MINIX is particularly inadequate for the relative ranking of carbohydrate conformers, as seen by the MAE of 4.41 kcal/mol for the *Carbohydrate* subset. The MAEs of more general subsets of conformational and deformation energies like *ANI1ccxCONF* and *MOLdef* are also high compared to other similar subsets, with values of 5.93 and 3.90 kcal/mol, respectively. Interestingly, HF/MINIX, despite its deficiencies, predicts a low MAE of 0.83 kcal/mol for the *Water-BE1*, which contains binding energies of small water clusters. As we shall see below, this surprisingly low MAE is likely the result of a fortuitous error cancellation. Finally, since all the MAEs presented in Figure 1 for HF/MINIX are relative to high-level reference data, the blue bars quantify the inaccuracy resulting from the shortcomings of HF/MINIX, primarily the absence of correlation in HF, including dispersion interactions, and the incompleteness of the MINIX basis set.

The D3 correction is parameterized to calculate the dispersion contribution to the total energy. This is expected to partially mitigate the absence of correlation in HF, particularly in the description of NCIs. Figure 1 shows that D3 applied to HF/MINIX causes a reduction in MAEs of most NCI types (by a factor of 1.3 to 3.0) including those with dominant electrostatic and dispersion character. For example, the application of D3 improves not only the non-covalent binding energies associated with hydrogen and halogen bonding interactions but also aliphatic-aliphatic interactions, stacked π-π interactions, non-stacked π-π interactions, and H-π interactions. Nevertheless, in some NCI types, D3 also appears to increase the MAEs associated with non-covalent binding energies. For example, D3 increases the MAEs of HF/MINIX for hydrogen-bonded small water clusters (*Water-BE1* subset) from 0.83 to 1.78 kcal/mol and of anion-containing complexes (*Anionic-BE1* subset) from 3.75 to 5.08 kcal/mol. This indicates that, for any interaction type which are already over-bound by HF/MINIX, application of D3 will lead to further overestimation in the binding energies, resulting in higher MAEs. Compared to non-covalent binding energy subsets, the degree of improvement from the application of the D3 correction is not as large for the conformational and deformation energy subsets. In fact, a contrasting performance is observed between some subsets. For example, D3 improves the MAEs of subsets representing protein and RNA conformers by a factor of about 1.3 but at the same time it deteriorates the MAEs of subsets representing carbohydrate, DNA, and hydrocarbon conformers by about 20% to 60%. It is important to point out that D3 was parametrized for use with basis sets larger than MINIX. The results confirm that D3 alone is an inadequate correction for small-basis-set HF, and further steps must be taken to mitigate the other errors associated with basis-set incompleteness from the underlying method.

HF-3c improves upon HF/MINIX by including the gCP and SRB correction terms, in addition to D3 with refitted parameters. Figure 1 shows that HF-3c reduces the MAEs of all subsets in the training set (compared to HF/MINIX) by a factor of 1.1 to 13.2, except the *Hydrocarbon-REL* subset of hydrocarbon conformational energies. For all non-covalent binding energy subsets (except the *Anionic-BE1* subset of anionic interactions), the HF-3c MAEs are below 1.43 kcal/mol. For the *Anionic-BE1* subset, the HF-3c MAE of 3.57 kcal/mol is only 0.18 kcal/mol lower than HF/MINIX. This shows that NCIs in anionic systems cannot be described adequately with HF-3c or uncorrected minimal-basis set HF. We observe from Figure 1 that HF-3c improves non-covalent binding energies more than it improves conformational and deformation energies. The improvement is also greatest where the errors from HF/MINIX are largest. For example, with HF-3c, the MAEs of some non-covalent binding energy subsets of biological relevance like *Carbohydrate-Protein* and *Nucleotide-Nucleotide* decrease from 7.10 and 7.83 kcal/mol to 0.54 and 1.01 kcal/mol, respectively. In contrast, the MAE of *Water-BE1* subset of hydrogen-bonded small water clusters decreases only from 0.83 to 0.72 kcal/mol.

Finally, we turn our attention to the HF/MINIX-ACP approach. From Figure 1, we see that HF/MINIX-ACP shows, in general, good performance for the training set. This, of course, is expected since the ACPs were developed using the data from this training set. Compared to HF/MINIX, the application of ACPs lead to an improvement factor in MAEs of 1.3 to 15.9. HF/MINIX-ACP predicts MAEs below 1.16 kcal/mol for all non-covalent binding energy subsets except the *Anionic-BE1* subset of anionic interactions, for which it is 2.96 kcal/mol. HF/MINIX-ACP shows a reduction in HF/MINIX MAEs by more than a factor of 5 for dispersion dominant NCI types like π-π interactions, H-π interactions, and aliphatic-aliphatic interactions (*π-stacking*, *Nucleotide-Nucleotide*, *Carbohydrate-Protein*, and *Hydrocarbon-BE* subsets). The results of HF/MINIX-ACP also show that a similar improvement in MAEs is obtained for interactions of mixed character (*Mixed NCIs* subset), interactions involving protein fragments (*Protein-Protein* subsets), and interactions involving gas adsorption on various substrates (*Gas-adsorption* subset). For other non-covalent binding energy subsets mainly of dominant electrostatic nature (*Hydrogen-bonding*, *Halogen-bonding*, and *Other-NCI* subsets) or some specific biological interactions (*Nucleotide-Protein* and *Biomolecule-Drug* subsets), ACPs lead to a reduction in the MAEs of HF/MINIX by a factor of almost 3.0 to 3.5. ACPs also reduce the MAEs of all conformational and deformation energy subsets, with an improvement factor ranging between 1.3 to 4.8. The most notable reduction is observed for conformers of biological relevance (*Protein*, *RNA*, and *Carbohydrate* subsets) and the diverse *ANI1cxxCONF* subset.

Among HF/MINIX, HF-D3/MINIX, HF-3c, and HF/MINIX-ACP methods, the overall best performers for the training set are HF-3c and HF/MINIX-ACP. The NCI types where HF-3c is the best performer include aliphatic-aliphatic interactions (*Hydrocarbon-BE* subset), π-π interactions (*π-stacking* subset), interactions of drugs to biomolecules (*Biomolecule-Drug* subset), interactions involving nucleotides and proteins (*Nucleotide-Protein* subset), and interactions involving gas adsorption on various substrates (*Gas-adsorption* subset). On the other hand, HF/MINIX-ACP shows the best performance for prediction of non-covalent binding energies in all other ten subsets of the training set. HF/MINIX-ACP also has the lowest MAEs among the compared methods for all the conformational and deformation energy subsets.

3.2 Performance of HF/MINIX-based methods on the validation set

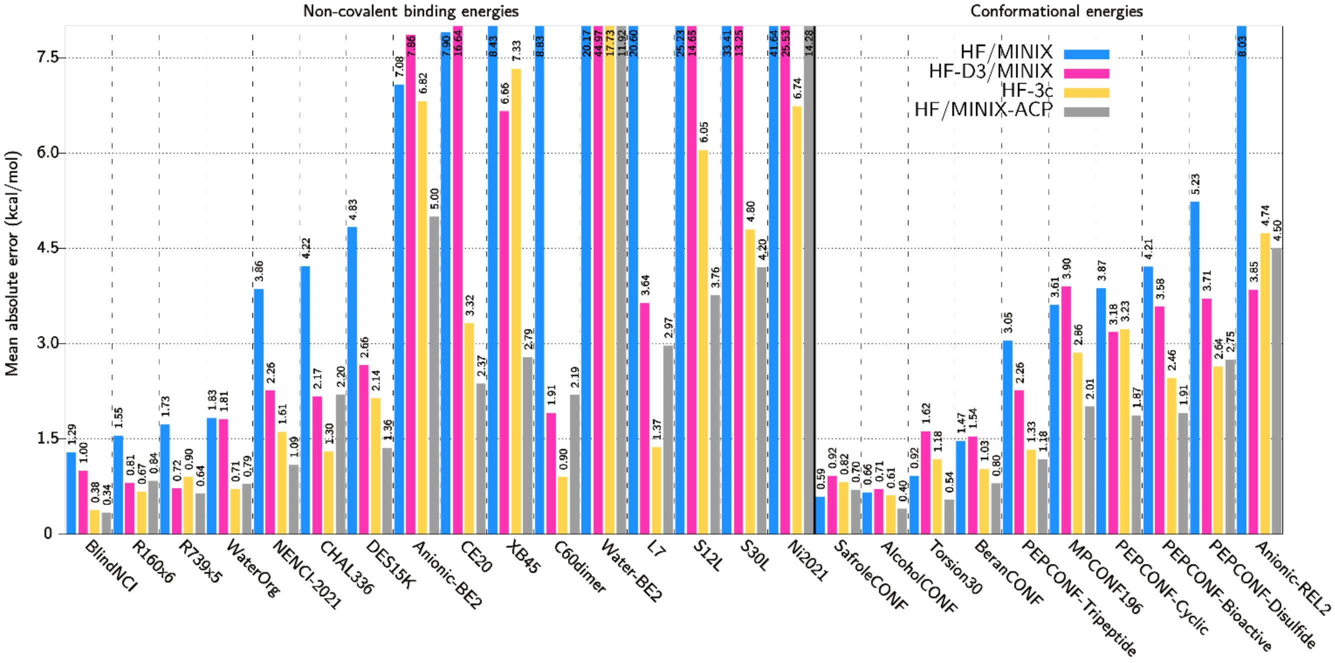


Figure 2. Mean absolute errors of HF/MINIX-based methods (relative to the reference data) for the validation set (Table 2). The methods shown are HF/MINIX (blue), HF-D3/MINIX (pink), HF-3c (yellow), and HF/MINIX-ACP (grey). The values for the mean absolute errors (in kcal/mol) are given atop the bars.

Let us now consider the results for the validation set. This allows comparing the performance of different methods against ACP-corrected methods in a more balanced way than by using the training set. The results for the validation set are presented in Figure 2. For the non-covalent binding energy subsets, the MAEs of HF/MINIX range from 1.29 to 41.64 kcal/mol. HF/MINIX yields undesirably large MAEs for not only interactions of mixed character found in biomolecules (*BlindNCI*, *DES15K*, *NENCI-2021*, *R160X6*, and *R739X5* subsets) but also for specific interactions like hydrogen bonding (*CE20*, *Water-BE2*, and *WaterOrg* subsets), halogen bonding (*XB45* subset), chalcogen bonding (*CHAL336* subset), and anionic interactions (*Anionic-BE2* subset). The larger MAEs are obtained for the subsets that contain relatively large complexes (*C60dimer*, *L7*, *S12L*, *S30L*, and *Ni2021* subsets). For the *Water-BE2* subset of hydrogen-bonded water clusters, the performance of HF/MINIX is severely hindered as a consequence of some negatively charged systems in the subset. As shown earlier, anionic systems are problematic for minimal-basis-set methods. The MAE for *Water-BE2* subset (20.17 kcal/mol) is more than an order of magnitude higher than that for *WaterOrg* (1.83 kcal/mol), a subset containing neutral hydrogen-bonded complexes of water clusters interacting with organic molecules. For similar reasons, HF/MINIX produces poor results for other subsets like *Anionic-BE2* (MAE=7.08 kcal/mol) and *Anionic-REL2* (MAE=8.03 kcal/mol) which also contain negatively charged species. The performance of HF/MINIX for non-covalently bound large complexes (*C60dimer, L7, S12L, S30L*, and *Ni2021* subsets) is also significantly worse (MAEs of 8.83 to 41.64 kcal/mol) than for other non-covalent interactions. Non-covalent binding energies of large complexes are challenging not only for HF/MINIX but also for many other electronic structure methods as most of these systems feature a combination of interaction types like hydrogen-bonding, halogen-bonding, π-π, H-π, ion-dipole, dispersion, etc. The reference energies for many of these systems are higher than 25 kcal/mol and go up to 416 kcal/mol, which is higher than other non-covalent binding energy subsets. Due to the large size of the systems, there are also some errors associated with the calculation of reference energies, because they are calculated using only approximate methods (like CIM-DLPNO-CCSD(T) for *Ni2021* and DLPNO-CEPA/1 for *C60dimer*) or back-corrected experimental data (like for *S12L* and *S30L)*. On the other hand, the MAEs of HF/MINIX for conformational and deformation energies present in the validation set, excluding the *Anionic-REL2* subset of anionic conformers, range from 0.59 to 5.23 kcal/mol. The MAEs for the subsets of small-molecule conformers (*AlcoholCONF*, *Torsion30*, and *BeranCONF* subsets) are large but reasonable (0.66–1.47 kcal/mol) compared to the MAEs (3.05–5.23 kcal/mol) for the subsets of proteinogenic conformers (*MPCONF196*, *PEPCONF-Tripeptide*, *PEPCONF-Cyclic*, *PEPCONF-Bioactive*, and *PEPCONF-Disulfide* subsets). A clear difference in performance is apparent between the subsets of small-molecule conformers and proteinogenic conformers: there is a difference of 1.58 kcal/mol between the highest MAE among small-molecule subsets (*BeranCONF*=1.47 kcal/mol) and the lowest MAE among proteinogenic subsets (*PECONF-Tripeptide*=3.05 kcal/mol). The poor overall results of HF/MINIX as discussed in this section, shown by the blue bars in Figure 2, clearly quantify the shortcomings of HF/MINIX. It demonstrates that the method without any additional correction is inadequate for modeling NCIs in large molecules as well as for conformer searching and structure prediction.

Application of the D3 correction to HF/MINIX lowers the MAEs of some subsets while increasing it for others. The notable cases where the MAEs of HF/MINIX do not change significantly or increase after applying D3 include the non-covalent binding energy subsets of hydrogen bonding interactions (*WaterOrg* and *CE20* subsets) and interactions involving anionic systems (*Water-BE2* and *Anionic-BE2* subsets). Similarly, in context of conformational energies, D3 does not improve the MAEs of HF/MINIX for all subsets of small-molecule conformers (*SafroleCONF*, *AlcoholCONF*, *Torsion30*, and *BeranCONF* subsets) as well as for the subset of macrocyclic peptide conformers (*MPCONF196* subset). However, for the remaining subsets in the validation set, D3 improves the MAEs over those of HF/MINIX by a factor of 1.2 to 5.7.

HF-3c performs significantly better compared to HF/MINIX with or without D3. The only two exceptions to this are small-molecule conformer subsets of *Torsion30* and *SafroleCONF*, where the MAEs of HF-3c are higher than HF/MINIX by about 28% and 39%, respectively. For all other subsets in the validation set, HF-3c improves the MAEs of HF/MINIX by a factor of 1.1 to 15.0, with the most significant improvement is observed for one of the subsets representing large molecules (*L7*, with a lowering of MAE from 20.60 to 1.37 kcal/mol). The results for HF-3c suggest that a significant reduction in error from HF/MINIX can be obtained once corrections are applied for both basis set incompleteness (using gCP and SRB terms) and absence of dispersion (using D3 with proper parameters for the MINIX basis set). However, it should also be noted that the 3c corrections to HF/MINIX can sometimes lead to higher MAEs than the D3 correction alone as seen for non-covalent binding energy subsets like *R739X5* and *XB45*, and for conformational energy subsets like *PEPCONF-Cyclic* and *Anionic-REL2*.

Finally, Figure 2 shows that applying ACPs to HF/MINIX reduces the MAEs of all subsets by a factor of 1.4 to 8.0, with the only exception being the small-molecule conformer subset of *SafroleCONF* (increase in MAE of 19%). The most significant reduction in MAE is seen for the large molecule non-covalent binding energy subset of *S30L*, where the MAE is lowered from 33.41 kcal/mol to 4.20 kcal/mol. For the *S30L* subset, even DFT with large basis sets fails to predict binding energies with such accuracy. For example, Brandenburg *et al.* reported an MAE of 6.6 kcal/mol using PBE-D3 at an estimated complete-basis-set limit.[27] In most subsets, the HF/MINIX-ACP predicted MAEs are lower than those predicted by either HF-D3/MINIX or HF-3c. A comparison of the MAEs of HF/MINIX-ACP on the training set and the validation set shows that parts of the validation and training sets that are similar have similar performance. For example, the good performance of HF/MINIX-ACP for interactions involving biomolecular systems (with MAEs ranging between 0.29–1.15 kcal/mol) in the training set transfers well to similar systems in the validation set, as seen by the MAEs ranging between 0.34 to 1.36 kcal/mol for non-covalent binding energy subsets like *BlindNCI*, *DES15K*, and *NENCI-2021*. Additionally, having trained the ACPs on many hydrogen-bonded systems helps reduce the MAE of *CE20* (a validation subset of hydrogen-bonded systems) from 7.90 to 2.37 kcal/mol. Similarly, the presence of halogen-bonded systems in the training set leads to a reduction in the MAE of *XB45* (a validation subset of halogen-bonded systems) from 8.43 to 2.79 kcal/mol. A comparison between the MAEs of HF/MINIX-ACP for the validation and training sets also allows us to see where the training set is incomplete. For example, the absence of training data for interaction energies of repulsive contacts in small organic molecules and chalcogen-bonding is reflected in the higher MAEs of HF/MINIX-ACP on *R160X6* and *CHAL336*, compared to HF-D3/MINIX (by about 1.5% to 3.5%) and to HF-3c (by about 20% to 41%).

Among HF/MINIX, HF-D3/MINIX, HF-3c, and HF/MINIX-ACP, the overall best performers for the validation set are HF-3c and HF/MINIX-ACP, same as for the training set. Among these two methods, HF-3c outperforms HF/MINIX-ACP for prediction of NCIs involving interactions of repulsive character (*R160X6* subset), hydrogen-bonded water and organic clusters (*WaterOrg* subset), chalcogen-bonded complexes (*CHAL336* subset), and some interactions between large complexes (*C60dimer*, *L7*, and *Ni2021* subsets). Conversely, HF/MINIX-ACP yields lower MAEs (0.34 to 5.00 kcal/mol) for all other ten subsets of various NCI types present in the validation set. For the conformational and deformation energy subsets, HF/MINIX-ACP yields the lowest MAEs (0.40 to 2.01 kcal/mol) for all subsets except *SafroleCONF* (HF/MINIX, 0.59 kcal/mol), *Anionic-REL2* and *PEPCONF-Disulfide* (HF-D3/MINIX and HF-3c, 3.85 and 2.64 kcal/mol).

3.3 Performance of HF/6-31G*-based methods

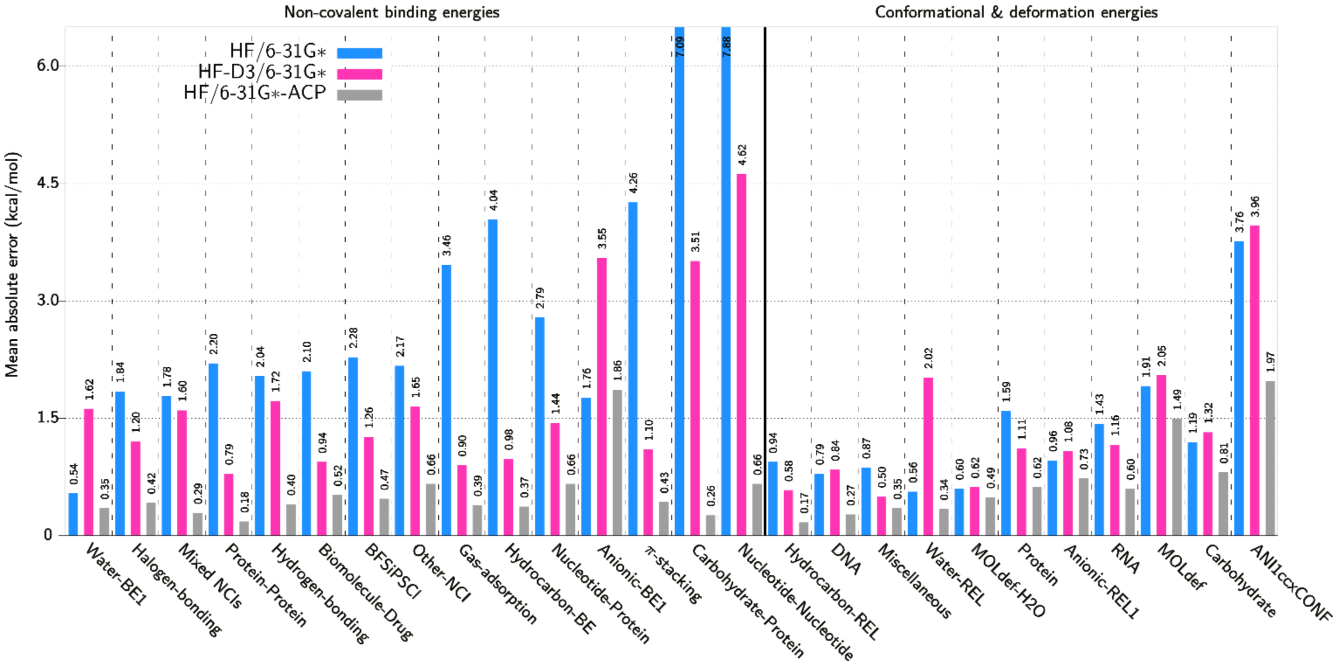


Figure 3. Mean absolute errors of HF/6-31G*-based methods (relative to the reference data) for the training set (Table 1). The methods shown are HF/6-31G* (blue), HF-D3/6-31G* (pink), and HF/6-31G*-ACP (grey). The values for the mean absolute errors (in kcal/mol) are given atop the bars.

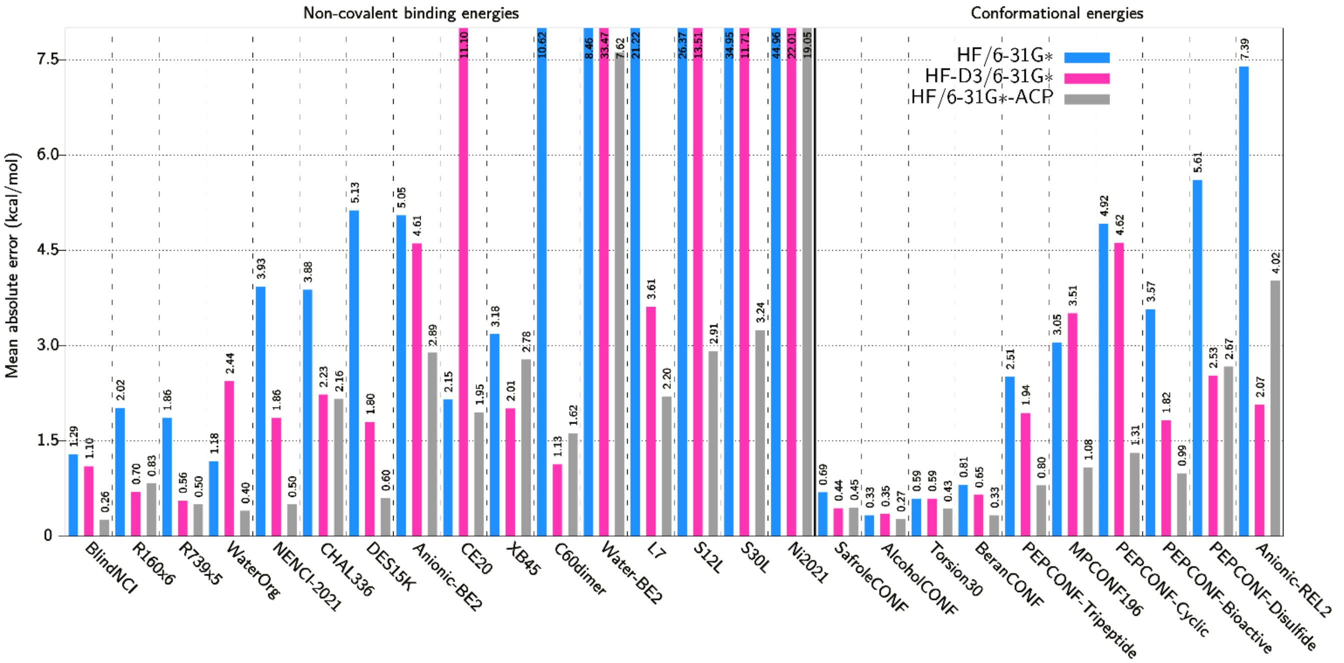


Figure 4. Mean absolute error of HF/6-31G*-based methods (relative to the reference data) for the validation set (Table 2). The methods shown are HF/6-31G* (blue), HF-D3/6-31G* (pink), and HF/6-31G*-ACP (grey). The values for the mean absolute errors (in kcal/mol) are given atop the bars.

Figures 3 and 4 show the results for HF/6-31G* methods on the training and validation sets, respectively. Overall, HF/6-31G* yields lower MAEs than HF/MINIX, but there are some exceptions. Comparison of HF/6-31G* and HF/MINIX for the non-covalent binding energy subsets in the training set (see Figures 1 and 3) shows that HF/6-31G* yields MAEs that are lower by about 1% to 53% than HF/MINIX for most subsets. The exceptions to this are subsets representing interactions of aliphatic-aliphatic character (*Hydrocarbon-BE* subset), π-π character (*π-stacking* subset), in proteins (*Protein-Protein* subset), nucleotides (*Nucleotide-Nucleotide* subset), and gas adsorption on substrates (*Gas-adsorption* subset). For the conformational and deformation energy subsets in the training set, the MAEs of HF/6-31G* are lower than HF/MINIX by about 13% to 73% for all subsets except those containing DNA-like and hydrocarbon-like conformers (*DNA* and *Hydrocarbon-REL* subsets). Similarly, a lowering in MAEs of HF/6-31G* compared to HF/MINIX by about 8% to 73% is also observed for many subsets in the validation set (see Figures 2 and 4). Therefore, these results confirm that going from a minimal basis set to a double-ζ basis set leads to an overall improvement due to smaller basis set incompleteness error.

Figures 3 and 4 show that the application of D3 leads to a general reduction in MAEs compared to uncorrected HF/6-31G* by a factor that ranges from 1.1 to 4.1 for the training set and 1.1 to 9.4 for the validation set. However, there are some subsets where the application of D3 increases the MAE by more than 100%. Such subsets include hydrogen-bonded systems (*Water-BE1*, *Water-REL* from the training set and *CE20*, *WaterOrg*, *Water-BE2* from the validation set) and anionic interactions (*Anionic-BE1* from the training set). The D3 correction to HF/6-31G* also results in lower MAEs than HF-D3/MINIX, but with some exceptions. Wherever the MAEs of HF-D3/6-31G* are lower than HF-D3/MINIX, the reduction varies between 2% to 75% for the training set and 1% to 70% for the validation set.

For HF/6-31G*-ACP, the same methodology used to develop ACPs for MINIX was applied to the 6-31G* basis set. Figures 3 and 4 show that the ACPs for HF/6-31G* consistently lower the MAEs of almost all subsets present in both the training and validation sets, to an even greater extent than those for HF/MINIX. The reduction in MAE brought by ACPs for all subsets of the training set (except the *Anionic-BE1* subset of anionic interactions) varies between a factor of 1.2 to 27.3. In the case of *Anionic-BE1*, the MAE of HF/6-31G* increases by about 6%. ACPs also reduce the MAEs of HF/6-31G* for all subsets of the validation set by a factor of 1.1 to 10.8. The MAEs of HF/6-31G*-ACP are also lower by about 10% to 60% compared to HF/MINIX-ACP for most subsets of the training and validation sets.

Among HF/6-31G*, HF-D3/6-31G*, and HF/6-31G*-ACP methods, the overall best performer for both the training and validation set is HF/6-31G*-ACP as its MAEs is lowest among the three methods for most subsets. The MAEs of HF-D3/6-31G* are only lower than both HF/6-31G* and HF/6-31G*-ACP for some subsets in the validation set, namely, those that contain systems representative of repulsive interactions (*R160X6* subset), halogen bonding interactions (*XB45* subset), dimers of C_{60} (*C60dimer* subset), and conformers of disulfide-bridged peptides and anionic systems (*PEPCONF-Disulfide* and *Anionic-REL2* subsets). It should be noted that the 3c correction scheme could not be directly assessed for use with HF/6-31G* because 3c was developed specifically for use with the MINIX basis set. However, we applied the contributions of 3c to HF/6-31G* with the aim of understanding its potential of transferability to 6-31G* basis set. The results associated with this test are presented in Figures S1 and S2 of the SI. We found out that the MAEs yielded after the addition of contributions from 3c to HF/6-31G* are lower than that of HF/6-31G* by about 13% to 91% for the training set and about 5% to 89% for the validation set. Interestingly, for some subsets, 3c performs even better in 6-31G* than in its own tailor-made basis set, MINIX, highlighting how significant the role of basis set incompleteness error is in the errors we are trying to mitigate. For example, in subsets of protein-like, RNA-like, and carbohydrate-like conformers (*Protein*, *RNA*, and *Carbohydrate* subsets from the training set), the respective MAEs are 1.64, 2.04, and 3.57 kcal/mol with HF-3c but 1.12, 0.97, and 1.97 kcal/mol after the addition of contributions from 3c to HF/6-31G*. Therefore, one can assume that if the 3c parameters were re-optimized for HF/6-31G*, a further improvement in most of the subsets might be obtained.

4. Summary and Outlook

In this work, we compared the accuracy of low-cost correction strategies designed to mitigate the underlying errors of minimal and double-ζ basis set based HF methods. For this purpose, two sets of data comprising 47,250 noncovalent binding energies and 58,630 relative conformation energies were assembled. We demonstrated how the accuracy of fast methods like HF/MINIX and HF/6-31G* can be significantly improved at predicting inter-and intra-molecular NCIs, i.e., non-covalent binding energies and conformational and deformation energies using correction strategies involving D3, 3c, and/or ACPs.

The D3 correction to HF/MINIX leads to a general reduction in MAEs produced by HF/MINIX for many of the data subsets. Both HF-3c and HF/MINIX-ACP are significantly better than D3 at reducing the underlying errors of HF/MINIX. Overall, HF/MINIX-ACP had the lowest MAEs for 20 of the 31 subsets of non-covalent binding energies. For conformational and deformation energies, HF/MINIX-ACP yielded the lowest MAEs for 18/21 subsets. Among the subsets of non-covalent binding energies, HF-3c yielded the lowest MAEs for the remaining 11 of the 31 subsets. It is important to underscore that ACPs were developed by fitting to about half the data points in the sets used for the performance comparisons. Nevertheless, the results indicate that ACPs used in conjunction with HF/MINIX can simultaneously correct some of the correlation missing from HF and for some of the basis set incompleteness error arising from the use of the MINIX basis set.

HF/6-31G* generally yielded MAEs that were lower (by about 1% to 73%) than HF/MINIX for most of the data subsets. This is in line with the expectation that double-ζ basis sets provide a better description of the systems under study than minimal basis sets, provided that basis set incompleteness does not dominate the non-covalent binding and conformational and deformation energies that comprise the datasets. The application of D3 to HF/6-31G* resulted in generally lower MAEs (about 1% to 75%) relative to HF-D3/MINIX. The application of ACPs developed for HF/6-31G* reduced MAEs by about 9% to 96% over HF/6-31G* alone, and by about 10% to 60% relative to ACP-corrected HF/MINIX.

The better performance of ACPs for HF/6-31G* than for HF/MINIX suggests that further improvements might be seen if ACPs were developed for even larger basis sets. However, there is a diminishing return associated with the use of ACPs for HF with larger basis sets as the calculation cost increases with the third power of the number of basis functions, thereby making the overall approach computationally expensive for large molecular systems. An alternative approach is to develop ACPs for use with small-basis-set HF-D3 or HF-3c. HF-3c produces results that are significantly better than

HF/MINIX, and the addition of HF-3c specific ACPs may reduce errors further. We intend to explore this idea in our future work.

The main drawback of any form of corrected, small-basis-set HF is that they cannot be applied far beyond the determination of the structure and non-covalent interactions. Therefore, these methods lack generality and, in particular, they lack the ability to predict thermochemical properties such as reaction energies and barrier heights. Grimme and co- workers have sought to address some of these issues by developing 3c corrections for use with larger basis sets and with density-functional theory methods, for example, PBEh with a double-ζ basis set (PBEh-3c)[58], HSE06 with a double-ζ basis set (HSE-3c)[59], B3LYP with a double-ζ basis set (B3LYP-3c)[60], B97-D with a triple-ζ basis set (B97-3c)[61], and r^2SCAN with a triple-ζ basis set (r^2SCAN-3c)[62]. Our future work will also involve developing ACPs for DFT-based methods with small and mid-sized basis sets with the goal of improving the prediction of both non-covalent and thermochemical properties.

In spite of this, methods that use corrections of the type explored in this work with small-basis-set HF have the potential to speed up the modeling of large molecular systems. For example, these low-cost methods can be used in cases where many preliminary structures need to be optimized and ranked by stability, after which more accurate methods can be used for refinement. Therefore, we believe that this analysis could be useful to the practitioners interested in such computational workflow.

References

(1) Biedermann F and Schneider H-J 2016 *Chem. Rev.* **116** 5216–300

(2) Hobza P and Muller-Dethlefs K 2009 *Non-Covalent Interactions: Theory and Experiment* (Cambridge: Royal Society of Chemistry)

(3) Scheiner S 2015 *Noncovalent Forces* (Springer International Publishing)

(4) Maharramov A M, Mahmudov K T, Kopylovich M N and Pombeiro A J L 2016 *Non-covalent Interactions in the Synthesis and Design of New Compounds* (Hoboken, NJ: John Wiley & Sons, Inc)

(5) Černý J and Hobza P 2007 *Phys. Chem. Chem. Phys.* **9** 5291–303

(6) Brini E, Fennell C J, Fernandez-Serra M, Hribar-Lee B, Lukšič M and Dill K A 2017 *Chem. Rev.* **117** 12385–414

(7) Mahadevi A S and Sastry G N 2016 *Chem. Rev.* **116** 2775–825

(8) Riley K E and Hobza P 2011 *WIREs Comput. Mol. Sci.* **1** 3–17

(9) Stone A 2013 *The Theory of Intermolecular Forces* 2nd edn (Oxford University Press)

(10) Otero-de-la-Roza A and DiLabio G A 2017 *Non-covalent Interactions in Quantum Chemistry and Physics : Theory and Applications* 1st edn (Elsevier Inc.)

(11) Kaplan I G 2006 *Intermolecular Interactions: Physical Picture, Computational Methods and Model Potentials* (Wiley)

(12) DiLabio G A and Otero-de-la-Roza A 2016 Noncovalent Interactions in Density Functional Theory *Reviews in Computational Chemistry* vol 29 (Wiley) 1–97

(13) Johnson E R, Mackie I D and DiLabio G A 2009 *J. Phys. Org. Chem.* **22** 1127–35

(14) Hermann J, DiStasio Jr R A and Tkatchenko A 2017 *Chem. Rev.* **117** 4714–58

(15) Stöhr M, Van Voorhis T and Tkatchenko A 2019 *Chem. Soc. Rev.* **48** 4118–54

(16) Gao H, Feng W, Li X, Li N, Du Y, Wu Y, Bai H and Qiao W 2019 *Phys. E Low-dimensional Syst. Nanostructures*

107 73–9

(17) Gao H, Sun Y, Zhang J, Wang Q, Wu Y and Bai H 2021 *Phys. E Low-dimensional Syst. Nanostructures* **127** 114532

(18) Hohenstein E G and Sherrill C D 2012 *WIREs Comput. Mol. Sci.* **2** 304–26

(19) Ratcliff L E, Mohr S, Huhs G, Deutsch T, Masella M and Genovese L 2017 *WIREs Comput. Mol. Sci.* **7** e1290

(20) Grimme S 2011 *WIREs Comput. Mol. Sci.* **1** 211–28

(21) Goerigk L 2017 A Comprehensive Overview of the DFT-D3 London-Dispersion Correction *Non-Covalent Interactions in Quantum Chemistry and Physics: Theory and Applications* (Elsevier Inc.) 195–219

(22) Grimme S, Antony J, Ehrlich S and Krieg H 2010 *J. Chem. Phys.* **132** 154104

(23) Johnson E R, Otero-de-la-Roza A, Dale S G and Dilabio G A 2013 *J. Chem. Phys.* **139** 214109

(24) Otero-de-la-Roza A and DiLabio G A 2017 *J. Chem. Theory Comput.* **13** 3505–24

(25) Otero-de-la-Roza A and Dilabio G A 2020 *J. Chem. Theory Comput.* **16** 4176–91

(26) Witte J, Neaton J B and Head-Gordon M 2016 *J. Chem. Phys.* **144** 194306

(27) Brandenburg J G, Hochheim M, Bredow T and Grimme S 2014 *J. Phys. Chem. Lett.* **5** 4275–84

(28) LeBlanc L M, Otero-de-la-Roza A and Johnson E R 2018 *J. Chem. Theory Comput.* **14** 2265–76

(29) Bannwarth C, Caldeweyher E, Ehlert S, Hansen A, Pracht P, Seibert J, Spicher S and Grimme S 2021 *WIREs Comput. Mol. Sci.* **11** e1493

(30) Caldeweyher E and Brandenburg J G 2018 *J. Phys. Condens. Matter* **30** 213001

(31) Christensen A S, Kubař T, Cui Q and Elstner M 2016 *Chem. Rev.* **116** 5301–37

(32) Yilmazer N D and Korth M 2015 *Comput. Struct. Biotechnol. J.* **13** 169–75

(33) Dral P O, Wu X, Spörkel L, Koslowski A, Weber W, Steiger R, Scholten M and Thiel W 2016 *J. Chem. Theory Comput.* **12** 1082–96

(34) Thiel W 2014 *WIREs Comput. Mol. Sci.* **4** 145–57

(35) Sure R and Grimme S 2013 *J. Comput. Chem.* **34** 1672–85

(36) Torres E and DiLabio G A 2013 *J. Chem. Theory Comput.* **9** 3342–9

(37) Mackie I D and DiLabio G A 2011 *Phys. Chem. Chem. Phys.* **13** 2780–7

(38) Holmes J D, Otero-de-la-Roza A and DiLabio G A 2017 *J. Chem. Theory Comput.* **13** 4205–15

(39) Prasad V K, Otero-de-la-Roza A and DiLabio G A 2018 *J. Chem. Theory Comput.* **14** 726–38

(40) van Santen J A and DiLabio G A 2015 *J. Phys. Chem. A* **119** 6703–13

(41) DiLabio G A and Koleini M 2014 *J. Chem. Phys.* **140** 18A542

(42) DiLabio G A, Koleini M and Torres E 2013 *Theor. Chem. Acc.* **132** 1389

(43) Torres E and DiLabio G A 2012 *J. Phys. Chem. Lett.* **3** 1738–44

(44) Mackie I D and DiLabio G A 2008 *J. Phys. Chem. A* **112** 10968–76

(45) DiLabio G A 2008 *Chem. Phys. Lett.* **455** 348–53

(46) Mackie I D and DiLabio G A 2010 *Phys. Chem. Chem. Phys.* **12** 6092

(47) Cao X and Dolg M 2011 *WIREs Comput. Mol. Sci.* **1** 200–10

(48) Dolg M and Cao X 2012 *Chem. Rev.* **112** 403–80

(49) Johnson E R and Becke A D 2006 *J. Chem. Phys.* **124** 174104

(50) Kruse H and Grimme S 2012 *J. Chem. Phys.* **136** 154101

(51) Frisch M J et al Gaussian 16, Revision B.01 (Gaussian, Inc., 2016)

(52) Neese F 2012 *WIREs Comput. Mol. Sci.* **2** 73–8

(53) Prasad V K, Otero-de-la-Roza A and DiLabio G A 2021 Fast and accurate quantum mechanical modeling of large molecular systems using atom-centered potentials developed for small-basis-set Hartree–Fock methods (in preparation)

(54) https://github.com/aoterodelaroza/dcp

(55) https://github.com/aoterodelaroza/acpfit

(56) https://github.com/aoterodelaroza

(57) Tibshirani R 2011 *J. R. Stat. Soc.: Series B (Statistical Methodology)* **73** 273–82

(58) Grimme S, Brandenburg J G, Bannwarth C and Hansen A 2015 *J. Chem. Phys.* **143** 054107

(59) Brandenburg J G, Caldeweyher E and Grimme S 2016 *Phys. Chem. Chem. Phys.* **18** 15519–23

(60) Pracht P, Grant D F and Grimme S 2020 *J. Chem. Theory Comput.* **16** 7044–60

(61) Brandenburg J G, Bannwarth C, Hansen A and Grimme S 2018 *J. Chem. Phys.* **148** 064104

(62) Grimme S, Hansen A, Ehlert S and Mewes J-M 2021 *J. Chem. Phys.* **154** 064103

(63) Jurečka P, Šponer J, Černý J and Hobza P 2006 *Phys. Chem. Chem. Phys.* **8** 1985–93

(64) Gráfová L, Pitoňák M, Řezáč J and Hobza P 2010 *J. Chem. Theory Comput.* **6** 2365–76

(65) Mintz B J and Parks J M 2012 *J. Phys. Chem. A* **116** 1086–92

(66) Řezáč J, Huang Y, Hobza P and Beran G J O 2015 *J. Chem. Theory Comput.* **11** 3065–79

(67) Černý J, Schneider B and Biedermannová L 2017 *Phys. Chem. Chem. Phys.* **19** 17094–102

(68) Řezáč J, Riley K E and Hobza P 2011 *J. Chem. Theory Comput.* **7** 2427–38

(69) Řezáč J, Riley K E and Hobza P 2011 *J. Chem. Theory Comput.* **7** 3466–70

(70) Witte J, Goldey M, Neaton J B and Head-Gordon M 2015 *J. Chem. Theory Comput.* **11** 1481–92

(71) Řezáč J, Dubecký M, Jurečka P and Hobza P 2015 *Phys. Chem. Chem. Phys.* **17** 19268–77

(72) Smith D G A, Burns L A, Patkowski K and Sherrill C D 2016 *J. Phys. Chem. Lett.* **7** 2197–203

(73) Burns L A, Vázquez-Mayagoitia Á, Sumpter B G and Sherrill C D 2011 *J. Chem. Phys.* **134** 084107

(74) Marshall M S, Burns L A and Sherrill C D 2011 *J. Chem. Phys.* **135** 194102

(75) Dilabio G A, Johnson E R and Otero-de-la-Roza A 2013 *Phys. Chem. Chem. Phys.* **15** 12821–8

(76) Burns L A, Faver J C, Zheng Z, Marshall M S, Smith D G A, Vanommeslaeghe K, MacKerell A D, Merz K M and Sherrill C D 2017 *J. Chem. Phys.* **147** 161727

(77) Kruse H, Banáš P and Šponer J 2019 *J. Chem. Theory Comput.* **15** 95–115

(78) Parker T M and Sherrill C D 2015 *J. Chem. Theory Comput.* **11** 4197–204

(79) Banáš P, Mládek A, Otyepka M, Zgarbová M, Jurečka P, Svozil D, Lankaš F and Šponer J 2012 *J. Chem. Theory Comput.* **8** 2448–60

(80) Kabeláč M, Valdes H, Sherer E C, Cramer C J and Hobza P 2007 *Phys. Chem. Chem. Phys.* **9** 5000–8

(81) Jakubec D, Hostaš J, Laskowski R A, Hobza P and Vondrášek J 2015 *J. Chem. Theory Comput.* **11** 1939–48

(82) Hostaš J, Jakubec D, Laskowski R A, Gnanasekaran R, Řezáč J, Vondrášek J and Hobza P 2015 *J. Chem. Theory Comput.* **11** 4086–92

(83) Jakubec D, Laskowski R A and Vondrasek J 2016 *PLoS One* **11** e0158704

(84) Stasyuk O A, Jakubec D, Vondrášek J and Hobza P 2017 *J. Chem. Theory Comput.* **13** 877–85

(85) Kozmon S, Matuška R, Spiwok V and Koča J 2011 *Chem. - A Eur. J.* **17** 5680–90

(86) Kozmon S, Matuška R, Spiwok V and Koča J 2011 *Phys. Chem. Chem. Phys.* **13** 14215–22

(87) Stanković I M, Blagojević Filipović J P and Zarić S D 2020 *Int. J. Biol. Macromol.* **157** 1–9

(88) Kumari M, Sunoj R B and Balaji P V 2012 *Org. Biomol. Chem.* **10** 4186–200

(89) Faver J C, Benson M L, He X, Roberts B P, Wang B, Marshall M S, Kennedy M R, Sherrill C D and Merz K M 2011 *J. Chem. Theory Comput.* **7** 790–7

(90) Kříž K and Řezáč J 2020 *J. Chem. Inf. Model.* **60** 1453–60

(91) Thanthiriwatte K S, Hohenstein E G, Burns L A and Sherrill C D 2011 *J. Chem. Theory Comput.* **7** 88–96

(92) Řezáč J, Fanfrlík J, Salahub D and Hobza P 2009 *J. Chem. Theory Comput.* **5** 1749–60

(93) Miriyala V M and Řezáč J 2017 *J. Comput. Chem.* **38** 688–97

(94) Řezáč J and Hobza P 2012 *J. Chem. Theory Comput.* **8** 141–51

(95) Řezáč J 2020 *J. Chem. Theory Comput.* **16** 2355–68

(96) Řezáč J 2020 *J. Chem. Theory Comput.* **16** 6305–16

(97) Sanders J M 2010 *J. Phys. Chem. A* **114** 9205–11

(98) Parrish R M and Sherrill C D 2014 *J. Am. Chem. Soc.* **136** 17386–9

(99) Steinmann S N and Corminboeuf C 2012 *J. Chem. Theory Comput.* **8** 4305–16

(100) Crittenden D L 2009 *J. Phys. Chem. A* **113** 1663–9

(101) Hill J G and Legon A C 2015 *Phys. Chem. Chem. Phys.* **17** 858–67

(102) Řezáč J, Riley K E and Hobza P 2012 *J. Chem. Theory Comput.* **8** 4285–92

(103) Setiawan D, Kraka E and Cremer D 2015 *J. Phys. Chem. A* **119** 1642–56

(104) Goerigk L, Hansen A, Bauer C, Ehrlich S, Najibi A and Grimme S 2017 *Phys. Chem. Chem. Phys.* **19** 32184–215

(105) Granatier J, Pitoňák M and Hobza P 2012 *J. Chem. Theory Comput.* **8** 2282–92

(106) Copeland K L and Tschumper G S 2012 *J. Chem. Theory Comput.* **8** 1646–56

(107) Smith D G A and Patkowski K 2014 *J. Phys. Chem. C* **118** 544–50

(108) Vogiatzis K D, Klopper W and Friedrich J 2015 *J. Chem. Theory Comput.* **11** 1574–84

(109) Smith D G A and Patkowski K 2015 *J. Phys. Chem. C* **119** 4934–48

(110) Li S, Smith D G A and Patkowski K 2015 *Phys. Chem. Chem. Phys.* **17** 16560–74

(111) Li W, Grimme S, Krieg H, Möllmann J and Zhang J 2012 *J. Phys. Chem. C* **116** 8865–71

(112) Temelso B, Archer K A and Shields G C 2011 *J. Phys. Chem. A* **115** 12034–46

(113) Mas E M, Bukowski R, Szalewicz K, Groenenboom G C, Wormer P E S and Van Der Avoird A 2000 *J. Chem. Phys.* **113** 6687–701

(114) Bukowski R, Szalewicz K, Groenenboom G C and van der Avoird A 2007 *Science* **315** 1249–52

(115) Bukowski R, Szalewicz K, Groenenboom G C and van der Avoird A 2008 *J. Chem. Phys.* **128** 094313

(116) Lao K U, Schäffer R, Jansen G and Herbert J M 2015 *J. Chem. Theory Comput.* **11** 2473–86

(117) Prasad V K, Otero-de-la-Roza A and DiLabio G A 2019 *Sci. Data* **6** 180310

(118) Goerigk L, Karton A, Martin J M L and Radom L 2013 *Phys. Chem. Chem. Phys.* **15** 7028

(119) Valdes H, Pluháčková K, Pitonák M, Řezáč J and Hobza P 2008 *Phys. Chem. Chem. Phys.* **10** 2747

(120) Kesharwani M K, Karton A and Martin J M L 2016 *J. Chem. Theory Comput.* **12** 444–54

(121) Mládek A, Krepl M, Svozil D, Čech P, Otyepka M, Banáš P, Zgarbová M, Jurečka P and Šponer J 2013 *Phys. Chem. Chem. Phys.* **15** 7295–310

(122) Mládek A, Banáš P, Jurečka P, Otyepka M, Zgarbová M and Šponer J 2014 *J. Chem. Theory Comput.* **10** 463–80

(123) Kruse H, Mladek A, Gkionis K, Hansen A, Grimme S and Sponer J 2015 *J. Chem. Theory Comput.* **11** 4972–91

(124) Csonka G I, French A D, Johnson G P and Stortz C A 2009 *J. Chem. Theory Comput.* **5** 679–92

(125) Chan B 2020 *J. Phys. Chem. A* **124** 582–90

(126) Sameera W M C and Pantazis D A 2012 *J. Chem. Theory Comput.* **8** 2630–45

(127) Marianski M, Supady A, Ingram T, Schneider M and Baldauf C 2016 *J. Chem. Theory Comput.* **12** 6157–68

(128) Gruzman D, Karton A and Martin J M L 2009 *J. Phys. Chem. A* **113** 11974–83

(129) Kozuch S, Bachrach S M and Martin J M L 2014 *J. Phys. Chem. A* **118** 293–303

(130) Martin J M L 2013 *J. Phys. Chem. A* **117** 3118–32

(131) Temelso B, Klein K L, Mabey J W, Pérez C, Pate B H, Kisiel Z and Shields G C 2018 *J. Chem. Theory Comput.* **14** 1141–53

(132) Fogueri U R, Kozuch S, Karton A and Martin J M L 2013 *J. Phys. Chem. A* **117** 2269–77

(133) Tahchieva D N, Bakowies D, Ramakrishnan R and Von Lilienfeld O A 2018 *J. Chem. Theory Comput.* **14** 4806–17

(134) Sharapa D I, Genaev A, Cavallo L and Minenkov Y 2018 *ChemPhysChem* **20** 92–102

(135) Sellers B D, James N C and Gobbi A 2017 *J. Chem. Inf. Model.* **57** 1265–75

(136) Folmsbee D and Hutchison G 2021 *Int. J. Quantum Chem.* **121** e26381

(137) Smith J S, Zubatyuk R, Nebgen B, Lubbers N, Barros K, Roitberg A E, Isayev O and Tretiak S 2020 *Sci. Data* **7** 1–10

(138) Smith B J, Swanton D J, Pople J A, Schaefer H F and Radom L 1990 *J. Chem. Phys.* **92** 1240–7

(139) Tschumper G S, Leininger M L, Hoffman B C, Valeev E F, Schaefer H F and Quack M 2002 *J. Chem. Phys.* **116** 690–701

(140) Taylor D E, Ángyán J G, Galli G, Zhang C, Gygi F, Hirao K, Song J W, Rahul K, Von Lilienfeld O A, Podeszwa R, Bulik I W, Henderson T M, Scuseria G E, Toulouse J, Peverati R, Truhlar D G and Szalewicz K 2016 *J. Chem. Phys.* **145** 124105

(141) Donchev A G, Taube A G, Decolvenaere E, Hargus C, McGibbon R T, Law K-H, Gregersen B A, Li J-L, Palmo K, Siva K, Bergdorf M, Klepeis J L and Shaw D E 2021 *Sci. Data* **8** 1–9

(142) Sparrow Z M, Ernst B G, Joo P T, Lao K U and DiStasio Jr R A 2021 arXiv:2102.02354

(143) Miriyala V M and Řezáč J 2018 *J. Phys. Chem. A* **122** 2801–8

(144) Kříž K, Nováček M and Řezáč J 2021 *J. Chem. Theory Comput.* **17** 1548–61

(145) Chan B, Gilbert A T B, Gill P M W and Radom L 2014 *J. Chem. Theory Comput.* **10** 3777–83

(146) Karton A, O'Reilly R J, Chan B and Radom L 2012 *J. Chem. Theory Comput.* **8** 3128–36

(147) Mehta N, Fellowes T, White J M and Goerigk L 2021 *J. Chem. Theory Comput.* **17** 2783–806

(148) Oliveira V, Kraka E and Cremer D 2016 T *Phys. Chem. Chem. Phys.* **18** 33031–46

(149) Romero-Montalvo E and DiLabio G A 2021 *J. Phys. Chem. A.* **125** 3369–77

(150) Bryantsev V S, Diallo M S, Van Duin A C T and Goddard W A 2009 *J. Chem. Theory Comput.* **5** 1016–26

(151) Anacker T and Friedrich J 2014 *J. Comput. Chem.* **35** 634–43

(152) Lao K U and Herbert J M 2013 *J. Chem. Phys.* **139** 034107

(153) Lao K U and Herbert J M 2015 *J. Phys. Chem. A* **119** 235–52

(154) Mardirossian N, Lambrecht D S, McCaslin L, Xantheas S S and Head-Gordon M 2013 *J. Chem. Theory Comput.* **9** 1368–80

(155) Sedlak R, Janowski T, Pitoňák M, Řezáč J, Pulay P and Hobza P 2013 *J. Chem. Theory Comput.* **9** 3364–74

(156) Calbo J, Ortí E, Sancho-García J C and Aragó J 2015 *J. Chem. Theory Comput.* **11** 932–9

(157) Risthaus T and Grimme S 2013 *J. Chem. Theory Comput.* **9** 1580–91

(158) Ambrosetti A, Alfè D, DiStasio Jr R A and Tkatchenko A 2014 *J. Phys. Chem. Lett.* **5** 849–55

(159) Sure R and Grimme S 2015 *J. Chem. Theory Comput.* **11** 3785–801

(160) Sharapa D I, Margraf J T, Hesselmann A and Clark T 2017 *J. Chem. Theory Comput.* **13** 274–85

(161) Ni Z, Guo Y, Neese F, Li W and Li S 2021 *J. Chem. Theory Comput.* **17** 756–66

(162) Mezei P D, Csonka G I, Ruzsinszky A and Sun J 2015 *J. Chem. Theory Comput.* **11** 360–71

(163) Zahn S, Macfarlane D R and Izgorodina E I 2013 *Phys. Chem. Chem. Phys.* **15** 13664–75

(164) Zhang H, Krupa J, Wierzejewska M and Biczysko M 2019 *Phys. Chem. Chem. Phys.* **21** 8352–64

(165) Kirschner K N, Heiden W and Reith D 2018 *ACS Omega* **3** 419–32

(166) Greenwell C and Beran G J O 2020 *Cryst. Growth Des.* **20** 4875–81

(167) Lahey S L J, Thien Phuc T N and Rowley C N 2020 *J. Chem. Inf. Model.* **60** 6258–68

(168) Řezáč J, Bím D, Gutten O and Rulíšek L 2018 *J. Chem. Theory Comput.* **14** 1254–66

Part V

The success of the various research studies conducted up to this point in the dissertation serves as the foundation for the ACP development work presented in this penultimate part. In the upcoming Chapters 8 and 9, new ACPs have been developed for use with minimal or double-ζ basis set Hartree–Fock (HF) or density-functional theory (DFT) methods. The ACP parametrization was performed on a large collection of high-quality reference data for the first time for ten elements commonly observed in organic chemistry and biochemistry (H, C, N, O, F, P, S, Cl) plus boron and silicon. The ACP parameters were obtained using the same regularized least-squares regression technique whose use was demonstrated to be beneficial in the proof-of-concept study. Besides, some or all the missing data sets generated in Part III of this thesis were also utilized to compile the appropriate training set for new ACP development along with other reference data collected from various sources in the literature. The total number of data points in the diverse training set amounts to 73,832 for ACPs with HF methods and 118,655 for ACPs with DFT methods.

Keeping in mind the target applications like fast geometry optimizations of large chemical structures, high-throughput conformer screening, and prediction of non-covalent interaction strengths in large systems, new ACPs were developed in Chapter 8 for being paired with small basis set HF methods, including those designed for use with the semi-empirical correction schemes tested in Chapter 7. In Chapter 9, we extended the applicability of ACPs beyond non-covalent properties and included thermochemical properties in the training set. In this case, the ACPs were designed for use with double-ζ DFT methods, which increased the target applications of ACPs to enable modeling chemical reactions involving large systems and fast transition state searches which was a limitation for the ACPs developed for HF methods.

The following two main research questions have been examined in the upcoming chapters: (i) how capable are the developed ACPs in reducing the errors in subsets of various molecular properties in the training set when applied to their corresponding small basis set HF or DFT methods? And (ii) are the developed ACPs robust enough to perform well for systems that were not used during the training process? For testing on systems outside the training set, an additional 32,048 data points for HF methods and 42,567 data points for DFT methods were used.

The primary data that support the findings of Chapters 8 and 9 are provided in Appendices 5 and 6 of this dissertation, respectively. Other supporting files have also been deposited to the figshare repository and are openly available at the following URL/DOI: https://doi.org/10.6084/m9.figshare.16912201.

Chapter 8

Fast and accurate quantum mechanical modeling of large molecular systems using small basis set Hartree–Fock methods corrected with atom-centered potentials

Abstract

There has been significant interest in developing fast and accurate quantum mechanical methods for modeling large molecular systems. In this work, by utilizing a machine-learning regression technique, we have developed new low-cost quantum mechanical approaches to model large molecular systems. The developed approaches rely on using one-electron Gaussian-type functions called atom-centered potentials (ACPs) to correct for the basis set incompleteness and the lack of correlation effects in the underlying minimal or small basis set Hartree-Fock (HF) methods. In particular, ACPs are proposed for ten elements common in organic and bio-organic chemistry (H, B, C, N, O, F, Si, P, S, and Cl) and four different base methods: two minimal basis sets (MINIs and MINIX) plus a double-ζ basis set (6-31G*) in combination with dispersion-corrected HF (HF-D3/MINIs, HF-D3/MINIX, HF-D3/6-31G*), and the HF-3c method. The new ACPs are trained on a very large set (73,832 data points) of non-covalent properties (interaction and conformational energies) and validated additionally on a set of 32,048 data points. All reference data is of complete basis set coupled-cluster quality, mostly CCSD(T)/CBS. The proposed ACP-corrected methods are shown to give errors in the tenths of a kcal/mol range for non-covalent interaction energies and up to 2 kcal/mol for molecular conformational energies. More importantly, the average errors are similar in the training and validation sets, confirming the robustness and applicability of these methods outside the boundaries of the training set. In addition, the performance of the new ACP-corrected methods is similar to complete basis set DFT but at a cost that is orders of magnitude lower, and the proposed ACPs can be used in any computational chemistry program that supports effective-core potentials without modification. It is also shown that ACPs improve the description of covalent and non-covalent bond geometries of the underlying methods and that the improvement brought about by the application of the ACPs is directly related to the number of atoms to which they are applied, allowing the treatment of systems containing some atoms for which ACPs are not available. Overall, the ACP-corrected methods proposed in this work constitute an alternative accurate, economical, and reliable quantum mechanical approach to describe the geometries, interaction energies, and conformational energies of systems with hundreds to thousands of atoms.

1. Introduction

Quantum mechanical (QM) methods are an indispensable tool for understanding chemical phenomena. When combined with a nearly complete basis set, high-level wavefunction theory methods can predict various thermochemical and structural properties with an accuracy comparable to, or even better than, experiments. However, such approaches have limited applicability because their computational cost increases steeply with the size of the system.[1–5] This precludes high-level wavefunction methods from being applied to study chemical and biological processes involving large molecular systems, such as enzymatic catalysis, protein folding, supra-molecular host-guest complexation, and many others.[6–14]

In the past few decades, a significant amount of effort has been devoted to developing efficient and accurate QM methodologies that can be applied to large molecular systems.[15–28] The application of QM modeling begins by selecting a set of approximations to solve the Schrödinger equation. One of the simplest and least expensive non-empirical approaches is the Hartree–Fock (HF) method. However, HF has a major shortcoming in that, by definition, it does not calculate any correlation energy, which results in overly repulsive dispersion interactions, bond lengths that are too short, and the poor prediction of various other molecular properties. In addition, the cost and accuracy of any QM method is strongly dependent on the choice of basis set, the set of functions used to describe the system's molecular orbitals. In HF, the computational cost scales roughly as the third power of the number of basis functions, with more sophisticated methods presenting an even steeper scaling. Calculations using either minimal or double-ζ basis sets are relatively inexpensive, but the use of these small basis sets introduces an additional error due to the insufficient number of basis functions. This basis set incompleteness error is severely detrimental to the method's accuracy. Therefore, even though minimal and double-ζ basis set HF offers a computationally inexpensive approach for modeling large molecular systems, a way needs to be devised to effectively mitigate the deleterious effect of missing electron correlation and basis set incompleteness error.[29,30]

Many existing semi-empirical QM methods are based on approximations to minimal basis set HF.[31–34] By construction, semi-empirical QM methods circumvent the calculation of certain two-electron integrals from the underlying minimal basis set HF approach while incorporating empirical parameters obtained by fitting to experimental or high-level theoretical reference data. These approximations substantially limit the accuracy of semi-empirical QM methods but in exchange reduce the computational cost below that of minimal basis set HF. Due to their reduced computational cost, semi-empirical QM methods have found extensive use in modeling large molecular systems. An example of a popular and more recent semi-empirical QM approach is the PM7 method of Stewart, which was also modified by Throssel and Frisch.[35,36]

Another approach that is similar in spirit to conventional semi-empirical QM methods is the HF-3c[37] method proposed by Sure and Grimme. The HF-3c method uses three separate geometry-dependent formulas[38,39] to add energy corrections ("3c") for the various deficiencies of minimal basis set HF: one to account for some of the missing dispersion interactions, and two to mitigate the effects of basis set incompleteness errors. Several other techniques have been proposed in the literature[40–53], reflecting the interest in developing computationally inexpensive methods for large systems.

Finding a good compromise between cost and accuracy is critical when modeling large molecular systems. HF-3c is an example of a QM method that strikes a good balance between these two desirable characteristics. Even though the cost of HF-3c is higher than most semi-empirical QM methods, it is still orders of magnitude cheaper than nearly complete basis set wavefunction theory or density-functional theory (DFT) based methods. On the other hand, the accuracy of HF-3c in describing molecular structures and non-covalent interaction strengths is similar to large basis set DFT.[54] These features allow HF-3c to be applied for fast geometry optimizations, conformer exploration, and prediction of non-covalent interaction energies in fairly large systems, with sizes between many hundreds and a few thousand atoms. This allows the QM description of (small) biological systems (proteins, nucleic acids, carbohydrates, lipids) as well as supramolecular host-guest complexes. The downside of HF-3c is that it is unable to accurately describe thermochemical quantities such as bond breaking and formation energies.[55] Grimme and co-workers have applied the 3c correction to a few density functional approximations to address this problem.[56–61]

Our previous works have shown that atom-centered potentials[62] (ACPs) offer a convenient means of improving the accuracy of HF and DFT based methods.[63–75] ACPs are one-electron potentials that share the same mathematical form as effective-core potentials[76,77] (ECPs) but do not replace any electrons. This allows ACPs to be used in most computational chemistry software packages without modifying the code. Additionally, ACPs are an economical way of mitigating the errors in the underlying methodology, since using them incurs only a small additional cost. In a previous proof-of-concept work, we developed a single set of ACPs for the H, C, N, and O elements to mitigate the shortcomings of dispersion-corrected minimal basis set HF.[63] The parameters for the ACPs were obtained by fitting to a set of 9,814 data points of non-covalent properties (interaction, conformational, and molecular deformation energies). In that work, we demonstrated the feasibility of the ACP correction approach by showing that ACPs developed for dispersion-corrected minimal basis set HF were able to accurately predict the mentioned non-covalent properties.

In this work, we build upon our previous study[63] and develop four ACP-corrected small basis set HF based methods. In all cases, the target applications are similar to those of HF-3c and our previous work,[63] namely structures and non-covalent interaction strengths. However, ACPs are developed for a larger set of atoms (H, B, C, N, O, F, Si, P, S, and Cl) than in our previous work, greatly increasing the applicability of the proposed methods. In addition, the use of the LASSO (Least Absolute Shrinkage and Selection Operator) regression[78–80] for fitting of ACP parameters greatly simplifies ACP development and allows using a training set about eight times as large and much more diverse than in earlier works,[63,67–69,72] resulting in more robust and more widely applicable ACPs. Three of the four new ACP-corrected methods are based on HF with minimal (MINIs[81] and MINIX[37]) or small double-ζ basis sets (6-31G*[82,83]) and use Grimme's D3[38,84,85] correction to account for the missing dispersion in HF. In addition, we also present a set of ACPs designed to improve the performance of the HF-3c method. In addition, our intention with the HF-3c-ACP method is to overcome the limitation imposed by the fact that ACPs are available only for the ten elements mentioned above. Since HF-3c parameters are available for most elements in the periodic table, we expect HF-3c-ACP to reduce to HF-3c performance for the atoms for which ACPs have not been developed, which in general should be in the minority. The newly developed ACP-corrected methods are assessed using an extensive validation set, demonstrating their performance and robustness.

2. Computational Methodology

2.1 Theoretical background

The procedure employed to develop the ACPs proposed in this work is similar to our earlier proof-of-concept study[63]. The mathematical form of an ACP is:

$$\hat{V}_{ACP} = \sum_{\alpha}\left(V_{local}^{\alpha}(r) + \sum_{l=0}^{L-1}\sum_{m=-l}^{l}\delta V_{l}^{\alpha}(r)\,|Y_{lm}\rangle\langle Y_{lm}|\right) \tag{1}$$

where $\delta V_l^{\alpha}(r) = V_l^{\alpha}(r) - V_{local}^{\alpha}(r)$, α represents the atoms on which the potentials are centered, and r is the distance to atom α. The $|Y_{lm}\rangle\langle Y_{lm}|$ represents projection operators using real spherical harmonics based on atom α with l angular quantum numbers and m magnetic quantum numbers. Equation 1 is the same general expression as ECPs[76,77]. The semi-local nature of the ACP arises from combining the first term (the *local* term), which only depends on the radial coordinate, with the second term (the *non-local* term), which incorporates the anisotropy via angular projections. The individual *local* and *non-local* terms in Equation 1 are represented by Gaussian-type functions:

$$V_{local}^{\alpha}(r) = \sum_{n=1}^{N} c_{local}^{\alpha} \exp(-\xi_{local}^{\alpha} r^2) \tag{2}$$

$$\delta V_l^{\alpha}(r) = \sum_{n=1}^{N} c_l^{\alpha} \exp(-\xi_l^{\alpha} r^2) \tag{3}$$

where the coefficients (c) and exponents (ξ) are adjustable parameters that are determined via a regularized least-squares fitting to reference data during ACP development (Section 2.2). The sum in Equations 2 and 3 runs over the total number (N) of Gaussian-type functions defined for atom α for the *local* and *non-local* potential terms. For ease of notation, we will represent the $V_{local}^{\alpha}(r)$ and $\delta V_l^{\alpha}(r)$ together as:

$$V_l^{\alpha}(r) = \sum_{n=1}^{N} c_{ln}^{\alpha} \exp(-\xi_{ln}^{\alpha} r^2) \qquad \text{for} \quad l = 0, 1, 2, \ldots, L \tag{4}$$

If the ACP operator (Equation 1) with the functional form of Equation 4 is treated as a perturbative correction to any Hamiltonian then the first-order perturbation energy correction induced by the ACPs is:

$$E_{ACP}(\{c_{ln}^{\alpha}\}, \{\xi_{ln}^{\alpha}\}) = \sum_{i} \langle \psi_i | \hat{V}_{ACP} | \psi_i \rangle \tag{5}$$

where the sum in Equations 5 runs over the occupied molecular orbitals. Substituting the expressions from Equations 1 and 4 into Equation 5 gives:

$$E_{ACP}(\{c_{ln}^{\alpha}\}, \{\xi_{ln}^{\alpha}\}) = \sum_{\alpha ln} c_{ln}^{\alpha} \sum_{i} \langle \psi_i | (|Y_{lm}\rangle \exp(-\xi_{ln}^{\alpha} r^2) \langle Y_{lm}|) | \psi_i \rangle \tag{6}$$

The $\Delta E_{ln}^{\alpha}(\xi_{ln}^{\alpha}) = \langle \psi_i | (|Y_{lm}\rangle \exp(-\xi_{ln}^{\alpha} r^2) \langle Y_{lm}|) | \psi_i \rangle$ integral, known as an ACP energy term, is the energy difference between the energy when an ACP with exponent ξ_{ln}^{α} is applied and the energy in absence of any ACP, divided by the ACP coefficient. Packing the coefficients c_{ln}^{α} and exponents ξ_{ln}^{α} into vectors and combining the terms in the inner sum of Equation 6 leads to:

$$E_{ACP}(\boldsymbol{c}, \boldsymbol{\xi}) = \sum_{\alpha ln} c_{ln}^{\alpha} \Delta E_{ln}^{\alpha}(\xi_{ln}^{\alpha}) = \boldsymbol{c} \cdot \mathbf{\Delta E}(\boldsymbol{\xi})^T \tag{7}$$

where $\mathbf{\Delta E}(\boldsymbol{\xi})^T$ is the vector of ACP energy terms. It should be noted that Equation 7 is only correct to first order in the ACP perturbation as the coefficients c_{ln}^{α} have an influence on the underlying wavefunction. Equation 7 becomes exact only in the limit of $\boldsymbol{c} \rightarrow \mathbf{0}$, and it is approximately correct if the ACP coefficients are small in magnitude. The deviation between the E_{ACP} obtained using a self-consistent calculation with the corresponding ACP and the linear estimate in Equation 7, which assumes the coefficients have no influence on the underlying wavefunction, is called the *non-linearity error*.

2.2 ACP development process

ACPs have features that make them useful to develop energy corrections for a QM method (see Reference 68 for more details): (i) ACPs generate energy correction terms based on the molecular orbitals (Equation 5) and are wavefunction-dependent, which means they include information from the chemical environment and the electronic wavefunction, (ii) the angular projection operators in the potential (Equation 1) produces energy correction terms that are dependent upon the local anisotropic environment of a given atom, (iii) the exponential form of the ACPs (Equation 4) ensures that a given ACP produces a correction that decays exponentially with interatomic distances, (iv) ACPs can be used with any software that uses Gaussian-type basis sets and ECPs, and (v) the use of ACPs incurs only in a small computational cost.

The first stage of the ACP development process involves assembling a comprehensive and diverse training set of target molecular properties. The training set should ideally consist of model systems composed of atoms for which ACPs are being developed. The training set should also contain molecules representing various chemical environments to ensure that the developed ACPs can be applied to diverse chemical systems. Since the focus of this work is to correct the deficiencies of small basis set HF based methods regarding molecular structures and non-covalent interactions, our training set contains data points of non-covalent interaction energies, molecular conformational energies, and molecular deformation energies.

Next, the exponents (ξ_{ln}^{α}), atoms (α), and angular momentum channels (l) are chosen, and the corresponding ACP energy terms ($\Delta E_{ln}^{\alpha}(\xi_{ln}^{\alpha})$) are calculated. The ACP energy term evaluation process is carried out by first obtaining SCF energy for every training set entry and then evaluating the ACP terms post-SCF. This approach speeds up the ACP energy term evaluation process which is important given the size of the training set and the number of ACP terms. Once the ACP energy terms have been computed for each target method/basis-set, exponents, angular momenta, and systems in the training set, the optimal ACP coefficients c_{ln}^{α} are determined using a regularized least-squares fit subject to a constraint on the sum of the absolute values of the coefficients. This constraint limits the magnitude of the ACP contributions and ensures that the correction arising from the ACP does not lead to significant non-linearity error, which would lead to disagreements between the predictions of our linear model (Equation 7) and the results obtained from the application of the ACP in an actual self-consistent calculation.

The training set is organized into subsets, corresponding to different molecular properties and data sources from the literature. Each subset of the training set is assigned a weight in the fitting procedure. The weight of subset i is calculated in the same way as in our previous work[63]:

$$w_i = \frac{1}{M_i N_i} \tag{8}$$

where M_i is the average of the absolute value of the reference energies and N_i is the number of data points in subset i. These weights account for differences in reference data magnitude and number of points in the subsets. The error function minimized in the fit is the weighted root-mean-square-error ($wRMSE$):

$$wRMSE = \sqrt{\frac{\sum_i (w_i \sum_j^{N_i} (y_{ref,j}^i - y_{method,j}^i)^2)}{\sum_i N_i}} \tag{9}$$

where j are the data points in the ith subset, $y_{ref,j}^i$ are the high-level reference energies for system j in the ith subset, and $y_{method,j}^i$ are the energies of the underlying method for which the ACPs are being developed.

The LASSO (Least Absolute Shrinkage and Selection Operator) regression[78–80], commonly employed in statistics and machine learning, is used to carry out the regularized least-squares fit. In LASSO, the $wRMSE$ in Equation 9 is minimized subject to the condition that l_1-norm of the ACP coefficients does not exceed a certain bound chosen beforehand:

$$\|c\|_1 = \sum_i |c_i| \tag{10}$$

The LASSO method is used to limit the magnitude of the ACP coefficients. In addition, for a given constraint, LASSO automatically selects the best subset of ACP terms and discards the others, resulting in ACPs with fewer terms, which is beneficial because it curbs the computational cost of applying the ACPs.

2.3 Training and validation data sets

The training set (Table 1) comprises non-covalent interaction energies, molecular conformational energies, and molecular deformation energies. This choice of training set properties is justified by the potential target applications of small basis set HF based methods, namely fast geometry optimizations and non-covalent interaction strengths in large systems as well as high-throughput[86] screening of conformers in combination with conformer search techniques[87–89]. These applications are useful, for instance, when performing exhaustive conformational searches of macrocyclic drugs[90–95] and other pharmaceutical candidates[96], and studying biochemical processes like protein folding[97–99] and puckering of nucleotides[100,101].

A successful method for non-covalent interactions must be able to accurately describe diverse non-covalent interaction motifs, which means that the training set must contain some of this diversity. For instance, the importance of π-π interactions is well-known in medicinal chemistry[102], structural biology[103,104], and organic electronics[105]. Such interactions also contribute to the stabilization of DNA[106,107] and proteins[108], control the strength and specificity of drug-protein interactions[109], and help in the rational design of supramolecules[110–112]. Other types of non-covalent interactions are also important in practice. For example, Berka *et al.* reported that aliphatic-aliphatic (or hydrophobic) interactions between amino acid backbone chains are the most abundant in proteins, particularly in the hydrophobic active site.[113] Hydrophobic interactions also control the structure and properties of lipid bilayers[114,115] and self-assembled supramolecules[116]. On the other hand, hydrogen bonding is probably the most studied non-covalent interaction.[117,118] Several other stabilizing non-covalent interactions with potential chemical and biological applications have also been studied, including halogen bonding[216–219], pnicogen bonding[220–222], and anionic[119–122] interactions. Therefore, we designed our training set to contain representative candidates from the interaction types mentioned above. This also allows us to assess the strengths and weaknesses of the developed ACPs regarding each interaction type.

In most cases, the subsets of the training set, and the molecular geometries and reference data in them, were adopted from the literature. Occasionally, the reference energies were re-calculated at a higher-level to improve their quality. In each subset of the training set, data points involving molecules containing atoms other than H, B, C, N, O, F, Si, P, S, and Cl were excluded. A detailed list of all the subsets used for our ACP training set is given in Table S1 of the SI. The number of data points in each subset varies depending on the availability of benchmark data sets.

The training set used here is almost eight times larger than in our previous work[63]. In total, the training set comprises 73,832 data points (167,275 molecular geometries) calculated mainly with complete basis set wavefunction theory methods (Table S1 in SI). The total number of non-covalent interaction energies, molecular conformational energies, and molecular deformation energies are 19,439, 44,105, and 10,288, respectively. The most abundant type of non-covalent interaction in the training set is hydrogen bonding. The mixed non-covalent interactions subset is second in abundance and features a mix of all common interactions found in large molecular systems. The large size of the training set ensures that no overfitting occurs when the least-squares fit is carried out.

In order to evaluate the performance and robustness of the new ACPs, we also assembled a validation set (Table 2), different from the training set, by compiling additional data sets from the literature. A detailed list of all the data sets included in the validation set is provided in Table S2 of the SI. In total, the validation set consists of 32,047 high-level data points (92,161 molecular geometries) calculated mainly with complete basis set wavefunction theory. The validation set contains 27,811 non-covalent interaction energies and 4,237 molecular deformation energies.

The structures and reference energies of all data points in the training and validation sets are given in the SI. In addition, the subsets that comprise the training and validation sets, grouped into categories to facilitate the analysis of the results, are listed in Tables 1 and 2. It should be noted that this subset categorization is in no way an exhaustive representation of the various types of systems in the training or validation set.

Table 1. List of data sets, grouped by category, in the ACP training set.

Category	**Data set(s)**	**Data points**	**Reference energy range (kcal/mol)**	**Description**
Non-covalent interaction energies of molecular complexes[a]:				
π-stacking	Pisub[b,123,124], Pi29n[125], BzDC215[126], C2H4NT[127]	379	-18.30 to +10.33	non-stacked and stacked π-π interactions
Hydrophobic	ADIM6[38,128,129], HC12[130]	18	-5.60 to -1.30	aliphatic-aliphatic interactions
Pnicogen-bonding	PNICO23[128,131]	23	-10.97 to -0.64	pnicogen bonding interactions
Halogen-bonding	Hill18[132], X40x10[133]	238	-14.14 to +11.95	halogen bonding interactions
Hydrogen-bonding	HBC6[134,135], MiriyalaHB104[136,137], IonicHB[138], HB375x10[139], IHB100x10[139], HB300SPXx10[140], CARBH12[128]	6,409	-37.01 to +16.30	hydrogen bonding interactions
Mixed NCIs	S22x5[135,141,142], S66x8[143–145], S66a8[144], A21x12[3,146,147], NBC10ext[127,135,148–150], 3B-69-DIM[151], 3B-69-TRIM[151], HW30[152]	1,895	-35.76 to +9.34	mixed-character non-covalent interactions

Category	Data set(s)	Data points	Reference energy range (kcal/mol)	Description
Anionic[c]	SSI-anionic[153], WatAA-anionic[b,154], HSG-anionic[135,155], PLF547-anionic[156], IonicHB-anionic[138], IHB100x10-anionic[139], Ionic43-anionic[157]	1,509	-135.11 to +88.94	anionic interactions
Biomolecule-Biomolecule	BBI[153], SSI[153], NucTAA[b,c,158–161], CarbhydBz[162], CarbhydNaph[163], CarbhydAroAA[b,164], CarbhydAro[b,165], WatAA[b, 154], HSG[135,155], PLF547[156], JSCH[141], DNAstack[166], DNA2body[166], ACHC[167], BDNA[168], NucBTrimer[b,169]	4,756	-100.86 to +64.19	interactions present in various biomolecules
Gas-Ligand	CH4PAH[170,171], CO2MOF[172], CO2PAH[173], CO2NPHAC[174], BzGas[175]	876	-6.02 to +12.17	interactions between gas molecules and substrate
Water-Water	Water38[176], Water1888[127,177–179], Water-2body[d,67]	2,336	-92.89 to +5.10	hydrogen-bonded water dimers and $(H_2O)_n$ clusters where n=3–10
BFSiPSCl	B-set[b,64], F-set[b,64], Si-set[b,64], P-set[b,64], S-set[b,64], Cl-set[b,64], Sulfurx8[180]	1,000	-68.05 to +21.57	monomers containing B, F, Si, P, S, and Cl atoms
Molecular conformational energies[e]:				
Small molecule	37Conf8[181], DCONF[182], ICONF[128], MCONF[183], Torsion21[184], MolCONF[185], ANI1ccxCONF[f,186]	41,224	+0.01 to +50.00	various molecules representing pharmaceuticals, catalysts, synthetic precursors, industrial chemicals, and organic compounds
Negatively charged[g]	PEPCONF-Dipeptide-anionic[b,187], MolCONF-anionic[185]	254	-0.47 to +10.96	negatively charged molecules
Biomolecule	PEPCONF-Dipeptide[b,187], TPCONF[188], P76[189], YMPJ[190], SPS[191], rSPS[192], UpU46[193], SCONF[128,194], DSCONF[195], SacchCONF[196], CCONF[197]	2,082	-4.09 to +19.74	molecules representative of proteins, DNA, RNA, and carbohydrates
Hydrocarbon	ACONF[198], BCONF[199], PentCONF[200]	421	+0.14 to +16.66	hydrocarbon-like molecules
$(H_2O)_{11}$	Undecamer125[201]	124	+0.06 to +1.87	$(H_2O)_{11}$ clusters
Molecular deformation energies[h]:				
Deformation	MOLdef[a,64], MOLdef-H2O[d,202,203]	10,288	-3.43 to +49.38	various molecules deformed along their normal modes

[a] defined as the difference between the energy of the complex and the sum of the monomer energies. A negative interaction energy indicates the complex is more stable than the separated monomers.

[b] the reference data was recalculated in this work at the DLPNO-CCSD(T)/CBS level of theory (see SI for more details), at geometries reported in the literature.

[c] comprises non-covalently bound dimers where at least one of the monomers is negatively charged.

[d] the reference data was calculated in this work at CCSD(T)/CBS level using the same extrapolation method as in Reference 176.

[e] defined as the difference between the energy of a particular conformer and a lower-energy conformer of the same molecule.

[f] contains mostly conformational energies but also some molecular deformation energies.

[g] comprises negatively charged conformers.

[h] defined as the difference between the energy of a molecule deformed along a particular normal mode and the energy of the same molecule at equilibrium.

Table 2. List of data sets, grouped by category, in the ACP validation set.

Category	Data set(s)	Data points	Reference energy range (kcal/mol)	Description
Non-covalent interaction energies of molecular complexes[a]:				
Mixed NCIs	BlindNCI[204], DES15K[205], NENCI-2021[206]	17,413	-33.78 to +186.83	mixed character non-covalent interactions
Hydrogen-bonding	CE20[207,208], WaterOrg[209]	2,396	-46.58 to -10.76	hydrogen bonding interactions
Halogen-bonding	XB45[210]	33	-13.11 to -0.89	halogen bonding interactions
Chalcogen-bonding	CHAL336[211]	48	-30.85 to -1.57	chalcogen bonding interactions
Repulsive contacts	R160x6[212], R739x5[213]	5,290	-12.02 to +6.79	close contact interactions
Anionic[b]	HW6Cl-anionic[214,215], HW6F-anionic[214,215], FmH2O10-anionic[214,215], SW49Bind345-anionic[216], SW49Bind6-anionic[216], Anionpi-anionic[217], IL236-anionic[218], DES15K-anionic[205], NENCI-2021-anionic[206], CHAL336-anionic[211], XB45-anionic[210], S30L-anionic[219]	2,525	-171.42 to +66.15	anionic interactions
$(H_2O)_{20}$ cluster	H2O20Bind10[215]	10	-200.54 to -196.59	$(H_2O)_{20}$ clusters
C_{60} dimer	C60dimer[220]	14	-6.88 to +12.07	C_{60} dimers
Large molecule	L7[221,222], S12L[9,11,222], S30L[219], Ni2021[223]	54	-416.08 to -1.68	large molecules relevant in supramolecular chemistry and biochemistry
Molecular conformational energies[c]:				
Small molecule	SafroleCONF[224], AlcoholCONF[225], BeranCONF[226], Torsion30[d,227]	2,193	+0.001 to +12.50	Safrole or 5-(2-propenyl)-1,3-benzodioxol) molecule, small alcohol molecules, small organic molecules, and biaryl drug-like molecules
Proteinogenic	MPCONF196[e,228], PEPCONF-Tripeptide[f,187], PEPCONF-Disulfide[g,187], PEPCONF-Cyclic[g,187], PEPCONF-Bioactive[g,187]	1,874	-0.47 to +81.00	peptide-like molecules
Negatively charged[h]	PEPCONF-Disulfide-anionic[g,187], PEPCONF-Bioactive-anionic[g,187]	170	+0.17 to +33.79	negatively charged molecules

[a] defined as the difference between the energy of the complex and the sum of energy of the monomers. A negative interaction energy indicates the complex is more stable than the separated monomers.

[b] comprises non-covalently bound complexes with at least one negatively charged monomer.

[c] defined as the difference between the energy of a particular conformer and a lower-energy conformer of the same molecule.

[d] only 30 systems used; we could not find the rest systems mentioned in Reference 227 in the supporting information

[e] only macrocyclic peptides considered.

[f] only a subset from the PEPCONF[187] database for which reference data was recalculated at the DLPNO-CCSD(T)/CBS level of theory (see SI for more details).

[g] available reference data was calculated at LC-ωPBE-XDM/aug-cc-pVTZ level of theory.

[h] comprises negatively charged conformers.

2.4 Technical details

Three sets of ACPs were developed for HF-D3 in combination with the minimal basis set MINIs, MINIX, and the double-ζ basis set 6-31G*. An additional set of ACPs was developed for HF-3c, which uses the MINIX basis set. The MINIX basis set was proposed at the same time as HF-3c[37], and is equivalent to MINIs for the first row atoms (H, B, C, N, O, and F) but employs an extra *d* basis function for Si, P, S, and Cl. Angular momentum channels up to the maximum angular momentum of the valence orbital basis functions for each atom present in the chosen basis set were used for ACP development. The maximum angular momentum values were: *s* for H (MINIs, MINIX, 6-31G*), *p* for B, C, N, O, F, Si, P, S, Cl (MINIs), *p* for B, C, N, O, F (MINIX), *d* for Si, P, S, Cl (MINIX), and *d* for B, C, N, O, F, Si, P, S, Cl (6-31G*).

Twenty-nine ACP exponents (ξ_{ln}^{α}) were chosen, with values: 0.12 to 0.30 in 0.02 steps, 0.40 to 2.00 in 0.10 steps, and 2.50 to 3.00 in 0.50 steps. It should be noted that the choice of exponents is different than in our previous work[63]. A careful evaluation of the computational cost associated with the calculation of ACP energy term integrals (Equation 7) with low exponent functions ($0.01 < \xi_{ln}^{\alpha} < 0.11$) suggested that ACPs containing exponents lower than 0.12 can lead to a significant increase in calculation time (especially for large molecules) compared to methods where ACPs are not applied. Therefore, choosing exponents higher than or equal to 0.12 ensures that the computational overhead is limited to a 10–30% increase relative to the uncorrected method.

The combination of ten atoms and twenty-nine exponents along with the various angular momentum channels results in 841 ACP terms for MINIs, 957 ACP terms for MINIX, and 1,102 ACP terms for 6-31G*. For our training set of 73,832 data points (167,275 molecular geometries), ACP development required a total of 140,678,275 (HF- D3/MINIs), 160,082,175 (HF-D3/MINIX), and 184,337,050 (HF-D3/6-31G*) single-point energies. Combined with the self-consistent calculations used to evaluate the impact of non-linearity error (1,338,200) and the calculations on the validation set to evaluate the performance of the ACPs (737,288), the total number of calculations for this project is 487,172,988. This complexity required the development of specialized software which we briefly describe next.

The parameters for the D3 dispersion correction used in this work correspond to those for the HF/aug-cc-pVTZ method with Becke-Johnson damping: $s_6 = 1.0$, $s_8 = 0.9171$, $a_1(BJ) = 0.3385$, and $a_2(BJ) = 2.8830$ Å. It should be noted that these D3 parameters are very close to those used in HF-3c[37]. The ACP energy term evaluation and fitting processes were carried out using the *dcp*[229] and *acpfit*[230] packages available in our GitHub repository[231]. These programs automatize and collate all the data required for ACP development. For the LASSO regression, we used the local linearization plus active set method proposed by Osborne *et al.*[232] and implemented in *octave/MATLAB* by Schmidt[233,234]. All single-point energy calculations with minimal or double-ζ basis set HF-D3 methods were performed with the *Gaussian-16*[235] software package. The HF-3c single-point energy calculations were performed with the *ORCA*[236] software package. All the SCF single-point energy calculations on the training and validation sets were carried out with the default settings. The post-SCF calculations for the ACP energy term evaluations (Equations 6 and 7) were executed using non-SCF multistep *Gaussian-16* jobs.

Once all the ACP energy terms for a particular target method were successfully computed, they were passed to the LASSO fit, resulting in an optimal set of ACPs for that method with minimum $wRMSE$ for a constraint of 25.0 au on the l_1-norm of coefficients. The ACPs proposed in this work contain approximately 6–19 terms per atom, and they are designed to be paired with the specific method for which they were developed (i.e., HF-D3/MINIs, HF-D3/MINIX, HF-D3/6-31G*, or HF-3c), and are not transferable to other methods. The ACP coefficients and exponents for each method are provided in the SI. An example of the usage of ACPs in the Gaussian software is also given in the SI.

3. Results and Discussion

3.1 Performance of ACPs for the training set

The optimal ACPs paired with their respective methods (HF-D3/MINIs, HF-D3/MINIX, HF-D3/6-31G*, and HF-3c) were applied self-consistently on the entire training set. The resulting non-covalent interaction energies, molecular conformational energies, and molecular deformation energies are compared to the corresponding reference data in Figures 1 and 2. The strip charts represent the error distribution as vertical lines for each method. The mean signed errors (MSEs) (open circle) and the standard deviations (SDs) of the errors (horizontal black lines) are also represented. The mean absolute errors (MAEs) and percentage change in the MAEs upon the application of ACPs (%ΔMAE) for each method are listed on the right. A detailed breakdown of the errors for each method and subset can be found in Table S3 of the SI.

Table S3 of the SI also lists the deviation between the prediction of the ACP performance from our linear model (the LASSO fitting procedure) against the actual results from using the ACP in self-consistent calculations. This comparison, which measures the extent of non-linearity error, shows that the deviation between the MAEs predicted by the linear model and the self-consistent calculations is under 10% for most of the subsets of the training set, with only a few exceptions. This indicates that the $\boldsymbol{l_1}$-norm constraint imposed in the LASSO fit was effective in preventing excessive non-linearity error and that the linear model used to develop ACPs is a faithful representation of their eventual performance as a correction method. In the following, the results obtained from the self-consistent application of ACPs are discussed for the different molecular properties in the training set.

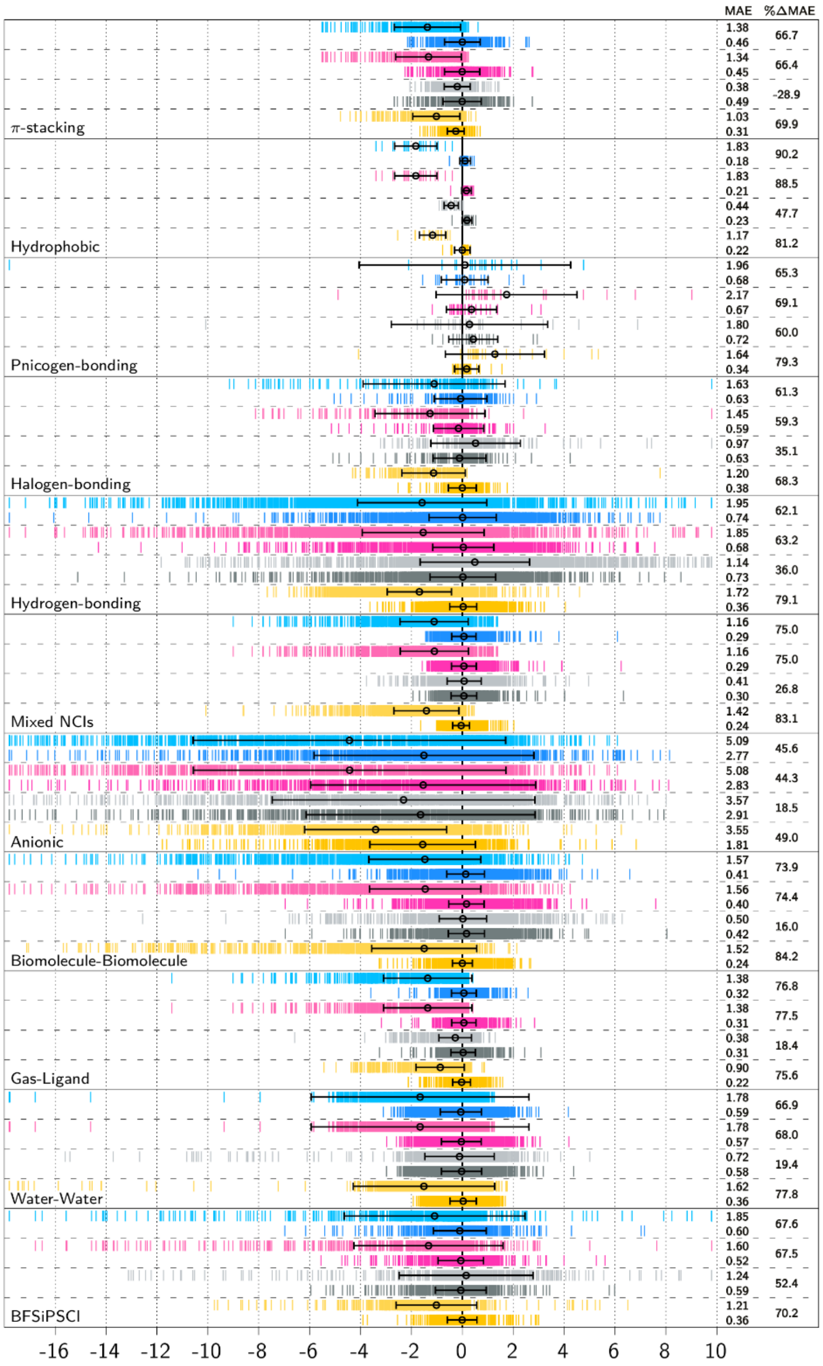


Figure 1. Error distribution (relative to the reference data, in kcal/mol) associated with non-covalent interaction energy subsets of the training set (see Table 1). Methods shown include HF-D3/MINIs (light blue), HF-D3/MINIs-ACP (blue), HF-D3/MINIX (light pink), HF-D3/MINIX-ACP (pink), HF-3c (light grey), HF-3c-ACP (grey), HF-D3/6-31G* (light yellow), and HF-D3/6-31G*-ACP (yellow). The black

circles represent the mean signed errors (MSEs, kcal/mol) and the black error bars are the standard deviations of the error (SDs, kcal/mol). The numbers on the right hand side of each panel are the mean absolute errors (MAEs, kcal/mol) and the percentage change in MAEs upon the application of ACPs (%ΔMAE) for each method. %ΔMAE is defined as [MAE(base method) – MAE(ACP-corrected method)] / MAE(base method) x 100%. The X-axis has been capped at -18 (left) and +10 kcal/mol (right) for clarity.

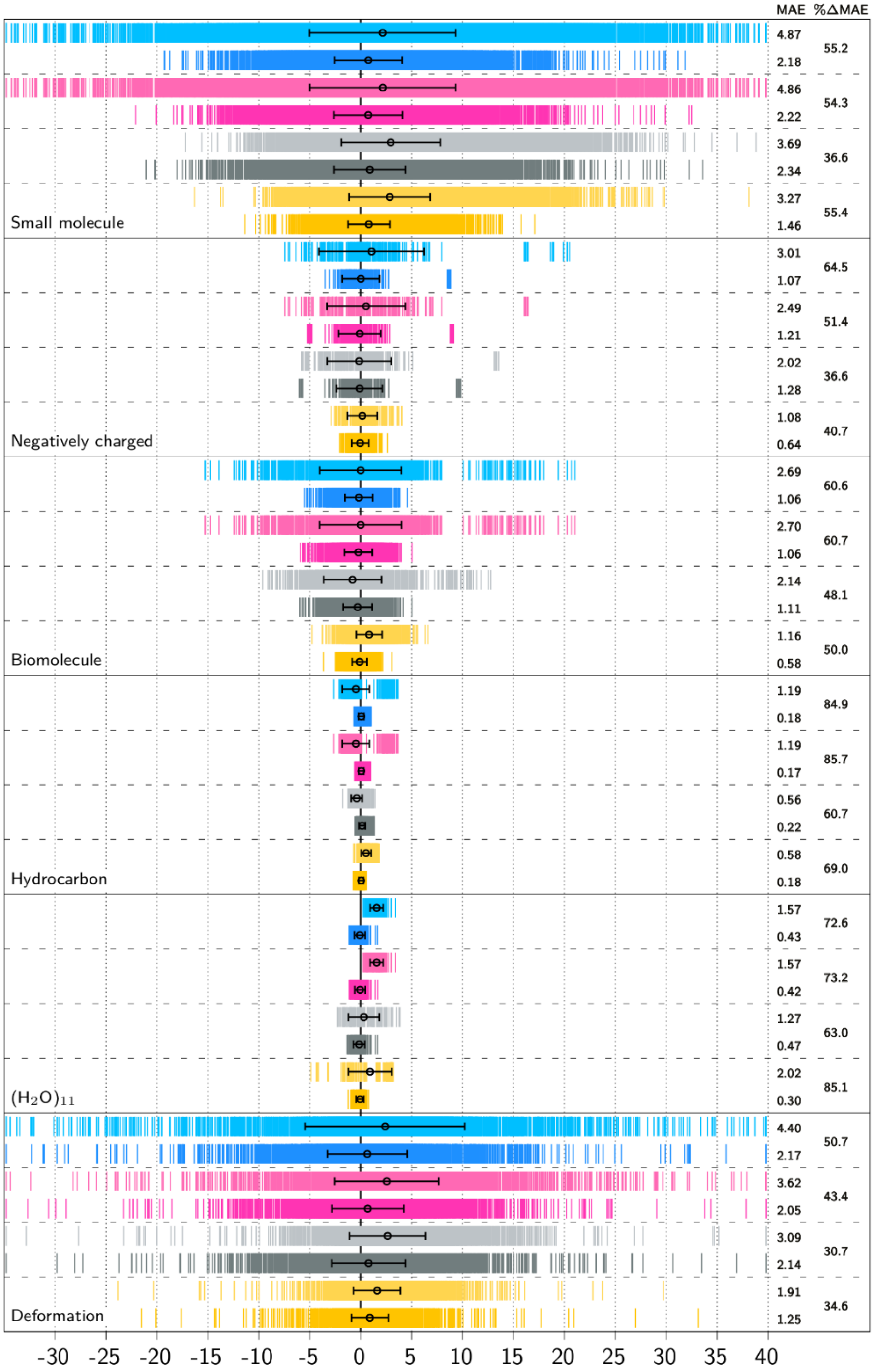


Figure 2. Error distribution (relative to the reference data, in kcal/mol) associated with molecular conformational and deformation energy subsets of the training set (see Table 1). Methods shown include HF-D3/MINIs (light blue), HF-D3/MINIs-ACP (blue), HF-D3/MINIX (light pink), HF-D3/MINIX-ACP

(pink), HF-3c (light grey), HF-3c-ACP (grey), HF-D3/6-31G* (light yellow), and HF-D3/6-31G*-ACP (yellow). The black circles represent the mean signed errors (MSEs, kcal/mol) and the black error bars are the standard deviations of the error (SDs, kcal/mol). The numbers on the right hand side of each panel are the mean absolute errors (MAEs, kcal/mol) and the percentage change in MAEs upon the application of ACPs (%ΔMAE) for each method. %ΔMAE is defined as [MAE(base method) – MAE(ACP-corrected method)] / MAE(base method) x 100%. The X-axis has been capped at -35 (left) and +40 kcal/mol (right) for clarity.

(i) Non-covalent interaction energies

The ACPs developed in this work have been trained on a wide range of non-covalent interaction types, including stacked and non-stacked π-π interactions, hydrophobic interactions, pnicogen bonding, halogen bonding, hydrogen bonding, and interactions of mixed and anionic nature. The proper description of each of these interactions is important for modeling large molecular systems, like proteins, where they operate co-operatively.[238] Figure 1 shows that the minimal or double-ζ basis set HF-D3 and HF-3c methods without ACPs have MAEs below 2 kcal/mol for different types of interactions except those of anionic nature (*Anionic* subset), where the MAEs are above 3.5 kcal/mol. Figure 1 also shows that HF-3c yields MAEs below 0.50 kcal/mol for *π-stacking*, *Hydrophobic*, and *Mixed NCIs* subsets, indicating that HF-3c is well suited to model systems that contain interactions of π-π, aliphatic-aliphatic, and mixed nature. The application of ACPs to minimal or double-ζ basis set HF-D3 and HF-3c methods mostly brings down the MAEs by about 44–90% (minimal basis set HF-D3), 49–84% (double-ζ basis set HF-D3), and 16–60% (HF-3c) for the range of interaction types covered in the subsets *π-stacking*, *Hydrophobic*, *Pnicogen-bonding*, *Halogen-bonding*, *Hydrogen-bonding*, *Mixed NCIs*, and *Anionic*.

Figure 1 shows that the application of ACPs to minimal or double-ζ basis set HF-D3 and HF-3c methods leads to an improved description of interactions of mixed, hydrogen bonding, and hydrophobic interaction types. π-π stacking interactions are also well described by our ACPs, except in the case of ACPs developed for HF-3c. Even though ACPs lead to a better description of anionic interactions than minimal or double-ζ basis set HF-D3 and HF-3c, the error spread is still large, evidencing the shortcomings of the underlying methods regarding anionic interactions.

Figure 1 also shows that the minimal or double-ζ basis set HF-D3 methods tend to over-estimate the interaction energies of almost all interaction types (except pnicogen bonding), with negative MSEs. On the other hand, the MSEs for HF-3c indicate that over-estimation or under-estimation of interaction energies depends on the nature of interaction type. The error spread in HF-3c is generally lower than minimal or double-ζ basis set HF-D3 methods, resulting in lower MAEs, MSEs, and SDs for this method.

When applied to minimal or double-ζ basis set HF-D3, ACPs correct the over-estimation in the interaction energies, resulting in a lower spread of errors and SDs, and a corresponding decrease in MSEs. Similarly, depending on the nature of interaction type, ACPs also improve the over-estimation or under-estimation tendencies of HF-3c for certain interaction types, causing a reduction in the corresponding MSEs.

We now examine the ACP performance for the more common interaction types in the training set, namely, hydrogen bonding and mixed interactions (*Hydrogen-bonding* and *Mixed NCIs* subsets). ACPs lower the MAEs of all four methods for the *Hydrogen-bonding* subset by about 63% (minimal basis set HF-D3), 79% (double-ζ basis set HF-D3), and 36% (HF-3c). For the *Mixed NCIs* subset, ACPs lower the MAEs by 75% (minimal basis set HF-D3), 83% (double-ζ basis set HF-D3), and 27% (HF-3c). It is also evident from Figure 1 that ACPs not only reduce the MAEs of the HF-D3 and HF-3c methods but also reduce the spread of errors and SDs and the bias. This is particularly true in the case of the HF-D3 methods. Some outliers with high error exist, which is natural given the very large size of the training set, but these errors are still lower than those predicted without ACPs. The individual errors for the *Mixed NCIs* subset are mostly within ±2 kcal/mol. Some of the systems with errors beyond ±2 kcal/mol are trimers with roughly twice the reference energies than the dimers in the training set. For the *Hydrogen-bonding* subset, an inspection of the errors beyond ±5 kcal/mol reveals that the ACPs over-stabilize the hydrogen bonding interactions of some complexes with polar bonds involving electronegative S, P, F, and Cl atoms.

The *π-stacking* subset with HF-3c-ACP is the only case where ACPs slightly increase the MAE of the base method (from 0.38 kcal/mol for HF-3c to 0.49 kcal/mol for HF-3c-ACP). However, it should be noted that for *π-stacking*, the ACPs developed for minimal or double-ζ HF-D3 methods, which initially have almost three times higher MAEs than HF-3c, do lead to a reduction in the MAEs by approximately 66–70%. A similar result occurs for the *Mixed NCIs* subset where MAEs of minimal or double-ζ HF-D3 methods are almost three times higher than HF-3c, and the application of ACPs reduce the MAEs for minimal or double-ζ HF-D3 methods by about 75–80% and by only about 27% for the HF-3c method. These two examples suggest that ACPs reduce the MAEs of the underlying methods when they are high and have a lesser impact on those subsets where the MAEs of the underlying method are already low. Consequently, in a few rare instances the performance of an underlying method with low initial MAE can be negatively, but only slightly, impacted by the application of ACPs. This is the case for HF-3c-ACP applied to π-π interactions.

A particular limitation of all methods in this work is the performance for the anionic systems in the *Anionic* subset. This subset is challenging for basis sets like MINIs, MINIX, and 6-31G* because of the lack of diffuse functions required to properly describe negatively charged species. Figure 1 shows that the error spread and the SDs of the *Anionic* subset are larger than other interaction types. Because the *Anionic* subset was used in the training set, ACPs improve the performance of all four methods for anionic interaction energies, with MAE reductions of 19–49%. However, there is obvious room for improvement, and it is likely that it can only be achieved by the inclusion of diffuse basis functions, which would incur in an additional computational cost.

An interesting observation from Figure 1 (also Table S3 of SI) is that when ACPs are developed for minimal basis set HF-D3 and HF-3c, the MAEs of the resulting ACP-corrected methods are very similar irrespective of whether the ACPs are applied to minimal basis set HF-D3 or HF-3c. For example, hydrogen bonding interactions (*Hydrogen-bonding* subset) with HF-D3/MINIs, HF-D3/MINIX, and HF-3c have MAEs of 1.95, 1.85, and 1.14 kcal/mol. The application of ACPs brings these MAEs down to very similar values (0.74, 0.68, and 0.73 kcal/mol) even though the MAEs for the uncorrected methods were quite different. Such consistency in the ACP-corrected MAEs is observed for most of the other types of interactions, and they indicate that the ACPs developed for minimal basis set HF-D3 are, to some extent, able to mitigate basis set incompleteness errors just like the gCP[37,39] and SRB[37] corrections of HF-3c. Also, since the ACPs developed for HF-3c in most cases improve on HF-3c, ACPs provide additional error mitigation beyond that offered by gCP[37,39] and SRB[37]. On the other hand, ACPs developed for double-ζ basis set HF-D3 result in lower MAEs than those used in combination with minimal basis set HF-D3 for each interaction type, indicating that systematic improvement can be obtained by using ACPs with larger basis sets. However, going beyond double-ζ would lead to a significant increase in computational cost, and would result in methods with limited applicability for large molecular systems.[239] In this regard, a better alternative would be the development of ACPs for use with double-ζ DFT methods, an idea that is currently being explored in our group.[240]

The performance of the proposed ACP based methods for the different interaction types and especially for the *Mixed NCIs* subset makes them promising for various applications. Keeping in mind our goal of designing low-cost approaches for modeling supramolecular and biological systems, we also assembled subsets and generated reference interaction energy data for prototypical non-covalently bound complexes relevant in biochemistry. Non-covalent interactions present in such systems are covered by the *Biomolecule-Biomolecule* subset. The *Biomolecule-Biomolecule* subset contains model systems representative of nucleotide-nucleotide interactions as well as protein fragments interacting with carbohydrates, nucleotides, drugs, water, and with other proteins. Such interactions are relevant in applications like protein folding[241,242], protein structure refinement[30,243], protein-ligand binding[244–246], intercalation[247], nucleobase stacking[248], and protein hydration[249], to name a few. Uncorrected minimal or double-ζ basis set HF-D3 in general overestimate the interaction energies in this subset, and application of the ACPs reduces this overestimation and decreases the error spread and SDs. Specifically, ACPs reduce the MAEs by about 74% (minimal basis set HF-D3) and 84% (double-ζ basis set HF-D3). HF-3c errors are centered around the zero-error average line with a relatively small spread, indicating that HF-3c is well suited for the complexes present in the *Biomolecule-Biomolecule* subset. Application of ACPs to HF-3c further reduces the MAE (by about 16%) except for a few systems: some nucleotide trimers and some nucleotide-amino acid complexes.

The *Gas-Ligand* subset comprises small molecules like CO_2, CH_4, and N_2 interacting with benzene, coronene, polycyclic aromatic hydrocarbons, polyheterocyclic aromatic compounds, and other functionalized organic molecules. The complexes present in the *Gas-Ligand* subset are representative of potential applications in the areas of chemical sensing, gas storage, and gas separation.[250–252] By training our ACPs to this subset we expect to extend their applicability to the modeling of gas adsorption on various porous materials.[253] The application of ACPs decreases the MAEs of all considered methods for the *Gas-Ligand* by about 77% (minimal basis set HF-D3), 76% (double-ζ basis set HF-D3), and 18% (HF-3c). The bias of minimal or double-ζ basis set HF-D3 and HF-3c towards over-estimating the interaction energies are also reduced with the application of ACPs, resulting in lower error spread and SDs.

The *Water-Water* subset contains interaction energies of water dimers at various intermolecular separations as well as small water clusters $(H_2O)_n$ with n=3–10. Potential target applications of ACPs trained against this subset are the modeling of aqueous environments, the study of surfaces of astrochemical interest[254–259], as well as performing *ab initio* molecular dynamics simulations of water[260–263]. The ACPs improve the MAEs of minimal or double-ζ basis set HF-D3 and HF-3c methods for *Water-Water* by about 68% (minimal basis set HF-D3), 78% (double-ζ basis set HF-D3), and 19% (HF-3c). Furthermore, Figure 1 shows that the large error spreads obtained with all underlying methods are significantly reduced by the ACPs, which bring the MSEs close to zero.

The last non-covalent interaction energy subset in the training set is *BFSiPSCl*, which contains complexes of monomers containing B, F, Si, P, S, and Cl. This subset extends the applicability of ACPs to systems like disulfide-linked proteins, covalent organic frameworks, functionalized silicon surfaces, and others. The application of ACPs results in a decrease in the MAEs of the *BFSiPSCl* subset by about 67% (minimal basis set HF-D3), 70% (double-ζ basis set HF-D3), and 52% (HF-3c), with a decrease in the error spread and SDs in all cases. The drop in MAE observed for the HF-3c-ACP is more significant for this subset than all other interaction energy subsets, and the reduction is also close to that observed in the *Pnicogen-bonding* subset, indicating that perhaps the HF-3c parametrization is not as good for these systems as for the more “usual” non-covalent interactions in the previous sets.

Finally, we consider a few illustrative examples for which we compare the performance of our ACP-corrected methods with some commonly used DFT methods in combination with large basis sets. For this purpose, we use representative data sets from Mardirossian and Head-Gordon's benchmarking work[237], for which nearly complete basis set DFT results have been reported in the literature. Specifically, we use the following sets: BzDC215[126] for π-π stacking interactions, HC12[130] for aliphatic-aliphatic interactions, S66x8[143–145] and 3B-69-DIM[151] for interactions of mixed nature, SSI[153] and HSG[135,155] for biomolecule-biomolecule interactions, Water38[176] for water-water interactions, and Sulfurx8[180] for interactions involving S atoms. The reported MAEs (in kcal/mol) of the DFT methods with the very large def2-QZVPDD basis set as well as the MAEs of the ACP-corrected methods are shown in Table 3. The table shows that the ACPs reduce the MAEs of minimal or double-ζ basis set HF-D3 and HF-3c methods in all the selected data sets and brings their MAE to a value close to or even lower than the almost complete basis set DFT methods. Therefore, Table 3 demonstrates that ACP-corrected methods have a performance similar to almost complete basis set DFT, but naturally at a cost that is reduced by orders of magnitude.

Table 3. Comparison of the mean absolute errors (MAEs) of various methods for selected data sets in the training set. (The MAEs lower than those calculated with various DFT methods using the def2-QZVPDD basis set are highlighted in bold.)

Data set[a]	**DFT functionals with def2-QZVPDD[b]**	**HF-D3/MINIs**	**HF-D3/MINIs-ACP**	**HF-D3/MINIX**	**HF-D3/MINIX-ACP**	**HF-3c**	**HF-3c-ACP**	**HF-D3/6-31G***	**HF-D3/6-31G*-ACP**
BzDC215[126]	0.41 [LC-ωPBE08-D3(BJ)]	0.92	**0.40**	0.85	**0.39**	**0.23**	0.44	1.27	**0.34**
HC12[130]	0.25 [M06-2X]	1.80	**0.20**	1.80	**0.22**	0.42	**0.25**	1.19	0.27
S66x8[143–145]	0.29 [CAM-B3LYP-D3(BJ)]	1.24	**0.25**	1.24	**0.26**	0.37	**0.27**	1.49	**0.23**
3B-69-DIM[151]	0.43 [M06-2X]	1.08	**0.42**	1.08	**0.41**	0.50	**0.42**	1.44	**0.25**
SSI[153]	0.17 [B3LYP-D3(BJ)]	0.87[c]	0.21[c]	0.87[c]	0.20[c]	0.28[c]	0.22[c]	0.76[c]	**0.15**[c]
HSG[135,155]	0.14 [B3LYP-D3(BJ)]	0.94[c]	0.18[c]	0.94[c]	0.19[c]	0.33[c]	0.19[c]	0.89[c]	**0.12**[c]
Water38[176]	2.85 [B3LYP-D3(BJ)]	30.62	**1.56**	30.62	**1.24**	7.67	**1.32**	19.51	**0.77**
Sulfurx8[180]	0.33 [BP86-D3(BJ)]	0.50	0.41	0.36	**0.29**	0.71	0.35	0.75	**0.19**

Data set[a]	DFT functionals with def2-QZVPDD[b]	HF-D3/MINIs	HF-D3/MINIs-ACP	HF-D3/MINIX	HF-D3/MINIX-ACP	HF-3c	HF-3c-ACP	HF-D3/6-31G*	HF-D3/6-31G*-ACP
YMPJ[190]	0.99 [B97-D]	1.73	**0.97**	1.74	1.00	2.32	1.04	**0.80**	**0.57**
SCONF[128,194]	0.57 [LC-ωPBE08-D3(BJ)]	5.20	0.59	5.20	**0.54**	1.47	**0.57**	1.57	0.64
BCONF[199]	0.34 [CAM-B3LYP-D3(BJ)]	2.40	**0.29**	2.40	**0.27**	0.58	**0.27**	1.25	**0.34**
PentCONF[200]	0.15 [B3LYP-D3(BJ)]	0.96	0.16	0.96	**0.15**	0.55	0.21	0.47	**0.15**

[a] details about the data sets can be found in Table S1 of the Supporting Information, [b] from Reference 237, [c] only non-negatively charged complexes

(ii) Molecular conformational energies

The purpose of the molecular conformational energy subsets of our training set is to inform the ACPs regarding how the potential energy surfaces of various molecules depend on the changes in rotatable bonds and torsional angles due to effects like π-conjugation, steric interactions, intramolecular hydrogen-bonding, and electron repulsion. Our *Small molecule* conformational energy subset is a good representative of such interactions that can be used to assess the performance of ACPs for conformational energies. The application of ACPs to the *Small molecule* subset leads to a reduction in the MAEs of about 55% (minimal and double-ζ HF-D3) and 37% (HF-3c), yielding MAEs ranging between 1.46–2.18 kcal/mol for ACPs with minimal or double-ζ basis set HF-D3 and 2.34 kcal/mol for ACPs with HF-3c. As seen in Figure 2, the spread of errors and SDs of the uncorrected methods is quite large: HF-D3/MINIs, for example, yields errors spanning -35 to +40 kcal/mol and an SD of 7.18 kcal/mol. The ACPs reduce the error spread of HF-D3/MINIs to about -20 to +30 kcal/mol and the SD to 3.31 kcal/mol. Similar observations can also be made for the HF-D3/MINIX, HF-D3/6-31G*, and HF-3c methods.

Conformers in the *Negatively charged* subset have an overall negative charge, which, as mentioned previously, is problematic for the minimal and double-ζ basis sets used in this work. Similar to the *Anionic* subset of non-covalent interaction energies, all uncorrected methods are inadequate for conformational energies of negatively charged species, which results in MAEs for the *Negatively charged* subset higher than for the other molecular conformational energy subsets. However, the application of ACPs yields relatively low MAEs (0.64–1.28 kcal/mol) compared to the uncorrected methods (1.08–3.01 kcal/mol), indicating that ACP-corrected methods are better suited to model molecular conformational energies of anionic systems.

Some other molecular conformational energy subsets used in the training set include the *Biomolecule*, *Hydrocarbon*, and *$(H_2O)_{11}$* subsets. The *Biomolecule* subset contains conformers of molecules that are biologically relevant, like proteins, DNA, RNA, and carbohydrates. The *Hydrocarbon* subset incorporates model systems of aliphatic nature relevant in lipids, polymers, fossil fuels, and organic chemistry. The *$(H_2O)_{11}$* contains systems relevant in the description of aqueous media.[264–266] The application of ACPs to the *Biomolecule*, *Hydrocarbon*, and *$(H_2O)_{11}$* subsets results in a significant drop in MAEs relative to the underlying methods, by about 61–86% (minimal basis set HF-D3), 50–85% (double-ζ basis set HF-D3), and 48–63% (HF-3c). Figure 2 shows that the error spread, SDs, and MSEs of minimal or double-ζ basis set HF-D3 and HF-3c methods are all reduced upon application of ACPs.

Same as for non-covalent interaction energies, Figure 2 shows that the application of ACPs brings down the MAEs of various molecular conformational energy subsets to similar values irrespective of whether the ACPs are applied to minimal basis set HF-D3 or HF-3c. For example, the MAEs of HF-D3/MINIs and HF-3c for the *Biomolecule* subset are 2.69 kcal/mol and 2.14 kcal/mol, respectively. Application of the corresponding ACPs results in a reduction of the MAEs to the very similar values of 1.06 kcal/mol and 1.11 kcal/mol. Like non-covalent interaction energies, the ACP for HF-D3/6-31G* yields lower MAEs compared to minimal basis set HF-D3 or HF-3c. For molecular conformational energies, the MAEs of the HF-3c-ACP method are notably lower (by about 31–61%) than that of HF-3c. Therefore, the ACPs developed for HF-3c offer a significant improvement beyond gCP[37,39] and SRB[37] for molecular conformation energies.

Finally, we compare the performance of our ACP-corrected methods relative to nearly complete basis set DFT results from the literature. For this, we consider a few representative data sets such as YMPJ[190] for amino acid conformers, SCONF[128,194] for carbohydrate-like conformers, BCONF[199] for butane-1,2-diol conformers, and PentCONF[200] for pentane conformers. The MAEs (in kcal/mol) of the DFT/def2-QZVPDD and the ACP-corrected methods for these data sets are shown in Table 3. Similar to non-covalent interaction energies, the application of ACPs reduces the MAEs of minimal or double-ζ basis set HF-D3 and HF-3c methods in all the selected data sets and are close to or even lower than the MAEs reported for the various functionals. Table 3 demonstrates that the proposed ACP-corrected methods are able to predict the conformational energies with an accuracy similar to large basis set DFT methods at a significantly lower computational cost.

(iii) Molecular deformation energies

The *Deformation* subset of the ACP training set contains energy differences between a molecule at its equilibrium geometry and the same molecule deformed along its various normal modes. Our intention with this subset is to improve the description of the molecular potential energy surfaces around the equilibrium geometries, and consequently improve the prediction of bond lengths and molecular geometries in general.

The fact that small basis set HF methods predict erroneous geometries is important for the study of large molecules like proteins, as discussed by Kulik *et al.*[267] and Schmitz *et al.*[268] Their findings suggest that small basis set HF methods without any correction give, in general, quite inaccurate protein structures. This is likely the reason why the HF-3c[37] method employs the semi-empirical SRB correction. In fact, the SRB correction itself was parametrized by fitting to the geometries of 107 small organic molecules computed at a higher level of theory.

The performance of our ACP-corrected methods for actual geometry optimizations is discussed in Section 3.4. The results for the *Deformation* subset in Figure 2 already suggest that ACPs improve the prediction of molecular geometries substantially. On application of ACPs, the MAEs of all four methods for this subset are reduced by about 35–51% (minimal and double-ζ HF-D3) and 31% (HF-3c). Figure 2 shows that even though the decrease in the spread of errors using ACPs is modest, the under-estimation in the prediction of molecular deformation energies of the underlying methods is greatly corrected by the ACPs, and the MSEs as well as the SDs decrease. Molecular deformations that are farthest from equilibrium have relatively high reference energies and result in errors higher than ±5 kcal/mol. Nevertheless, the application of ACPs predict individual errors that are lower than ±5 kcal/mol for 85% (or more) of the data points out of a total of 10,288.

3.2 Performance of ACPs for the validation set

The results regarding the application of ACPs to the systems in the validation set (Table 2) are presented in Figure 3. The figure includes the signed error distribution, MSEs, MAEs, and SDs of minimal or double-ζ basis set HF-D3 and HF-3c methods with and without ACPs, as well as the percentage change in MAEs upon application of ACPs (%ΔMAE) for each method. A detailed breakdown of the errors by method and subset can be found in Table S4 of the SI.

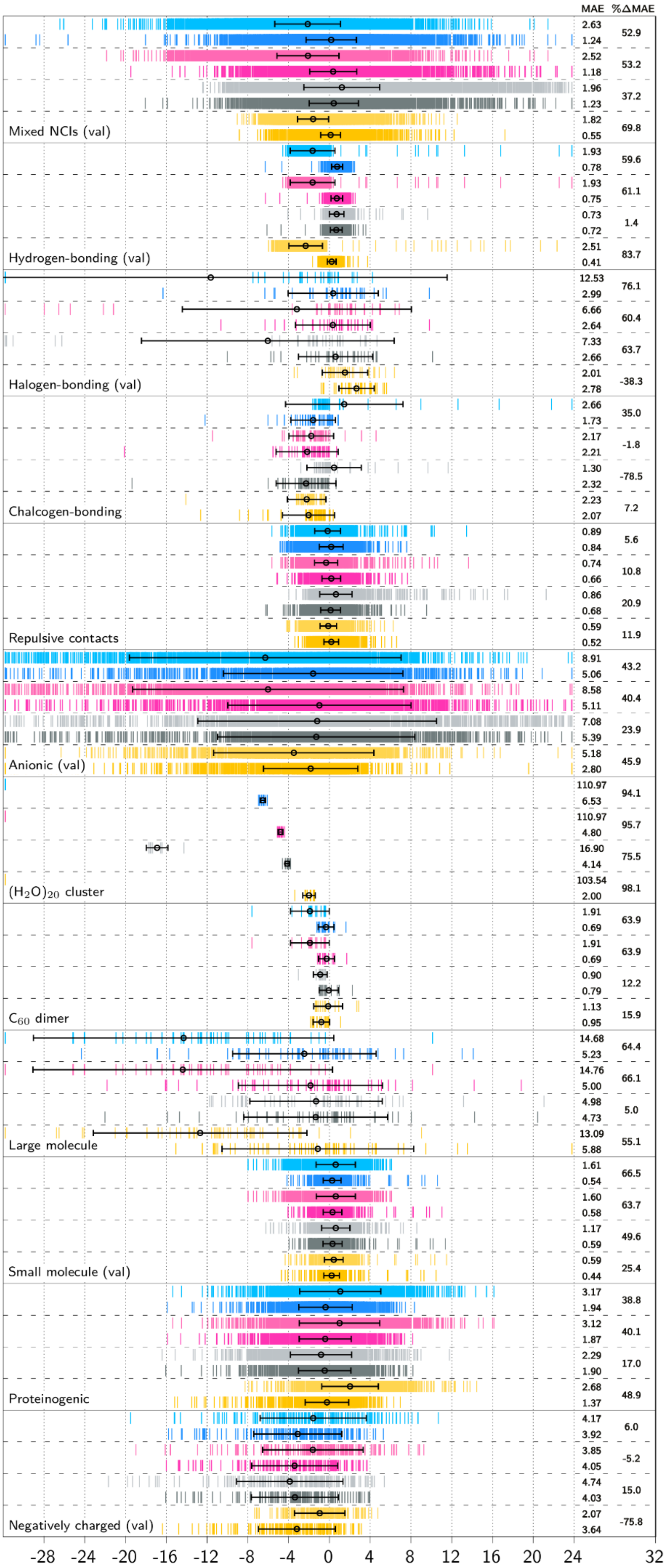

MAE
%ΔMAE
Mixed NCIs (val)
2.63
1.24
52.9
2.52
1.18
53.2
1.96
1.23
37.2
1.82
0.55
69.8
Hydrogen-bonding (val)
1.93
0.78
59.6
1.93
0.75
61.1
0.73
0.72
1.4
2.51
0.41
83.7
Halogen-bonding (val)
12.53
2.99
76.1
6.66
2.64
60.4
7.33
2.66
63.7
2.01
2.78
-38.3
Chalcogen-bonding
2.66
1.73
35.0
2.17
2.21
-1.8
1.30
2.32
-78.5
2.23
2.07
7.2
Repulsive contacts
0.89
0.84
5.6
0.74
0.66
10.8
0.86
0.68
20.9
0.59
0.52
11.9
Anionic (val)
8.91
5.06
43.2
8.58
5.11
40.4
7.08
5.39
23.9
5.18
2.80
45.9
$(H_2O)_{20}$ cluster
110.97
6.53
94.1
110.97
4.80
95.7
16.90
4.14
75.5
103.54
2.00
98.1
C_{60} dimer
1.91
0.69
63.9
1.91
0.69
63.9
0.90
0.79
12.2
1.13
0.95
15.9
Large molecule
14.68
5.23
64.4
14.76
5.00
66.1
4.98
4.73
5.0
13.09
5.88
55.1
Small molecule (val)
1.61
0.54
66.5
1.60
0.58
63.7
1.17
0.59
49.6
0.59
0.44
25.4
Proteinogenic
3.17
1.94
38.8
3.12
1.87
40.1
2.29
1.90
17.0
2.68
1.37
48.9
Negatively charged (val)
4.17
3.92
6.0
3.85
4.05
-5.2
4.74
4.03
15.0
2.07
3.64
-75.8
-28
-24
-20
-16
-12
-8
-4
0
4
8
12
16
20
24
32

Figure 3. Error distribution (relative to the reference data, kcal/mol) associated with the validation set (see Table 2). The top nine panels represent non-covalent interaction energy subsets while the bottom three panels represent molecular conformational energy subsets. Methods shown include HF-D3/MINIs (light blue), HF-D3/MINIs-ACP (blue), HF-D3/MINIX (light pink), HF-D3/MINIX-ACP (pink), HF-3c (light grey), HF-3c-ACP (grey), HF-D3/6-31G* (light yellow), and HF-D3/6-31G*-ACP (yellow). The black circles represent the mean signed errors (MSEs, kcal/mol) and the black error bars are the standard deviations of the error (SDs, kcal/mol). The numbers on the right hand side of each panel are the mean absolute errors (MAEs, kcal/mol) and the percentage change in MAEs upon the application of ACPs (%ΔMAE) for each method. %ΔMAE is defined as [MAE(base method) – MAE(ACP-corrected method)] / MAE(base method) x 100%. The X-axis has been capped at -32 (left) and +24 kcal/mol (right) for clarity. The black circles and error bars of HF-D3/MINIs, HF-D3/MINIX, and HF-D3/6-31G* methods for *$(H_2O)_{20}$ cluster* subset are absent from the figure due to MAEs being higher than 100 kcal/mol.

The results show that the when the ACPs are applied to the *Mixed NCIs* validation subset ("*Mixed NCIs (val)*"), the MAEs of all methods decrease, by 53% (minimal basis set HF-D3), 70% (double-ζ basis set HF-D3), and 37% (HF-3c). This reduction in MAEs is similar to what was observed for the mixed character interactions in the training set, confirming the robustness of the ACPs for non-covalent interactions when applied to systems outside the training set. One particular data set present in the *Mixed NCIs (val)* subset is BlindNCI[204]. Taylor *et al.*[204] reported an MAE of 0.34 kcal/mol with the M11/aug-cc-pVTZ method for the BlindNCI data set, which is almost equivalent to the MAE obtained with HF-D3/MINIs-ACP, HF-D3/MINIX-ACP, and HF-3c-ACP, and almost 28% higher than HF-D3/6-31G*-ACP (see Table 4). As in the training set, the performance of ACPs in the description of non-covalent interaction energies in the validation set is similar in quality to large basis set DFT methods.

Table 4. Comparison of the mean absolute errors (MAEs) of various methods for selected data sets in the validation set. (The MAEs that are lower than the DFT methods are highlighted in bold.)

Data set[a]	**DFT functional with a large basis set**	**HF-D3/MINIs**	**HF-D3/MINIs-ACP**	**HF-D3/MINIX**	**HF-D3/MINIX-ACP**	**HF-3c**	**HF-3c-ACP**	**HF-D3/6-31G***	**HF-D3/6-31G*-ACP**
BlindNCI[204]	0.34[b] [M11/aug-cc-pVTZ]	1.00	**0.34**	1.00	0.36	0.38	0.35	1.10	**0.25**
CE20[207,208]	1.72[c] [M06-2X/6-311+G(3df,2p)]	16.64	**1.63**	16.64	**1.55**	3.32	1.84	11.10	**1.31**

Data set[a]	DFT functional with a large basis set	HF-D3/MINIs	HF-D3/MINIs-ACP	HF-D3/MINIX	HF-D3/MINIX-ACP	HF-3c	HF-3c-ACP	HF-D3/6-31G*	HF-D3/6-31G*-ACP
WaterOrg[209]	0.44[d] [B3LYP-D3(BJ)/6-31+G**-BSIP]	1.81	0.77	1.81	0.75	0.71	0.71	2.44	**0.40**
CHAL336[211]	1.18[e] [BLYP-D3(BJ)/ma-def2-QZVPP]	2.66[l]	1.73[l]	2.17[l]	2.21[l]	1.30[l]	2.32[l]	2.23[l]	2.07[l]
H2O20Bind10[215]	8.84[f] [B3LYP-D3(BJ)/def2-QZVPPD]	110.97	**6.53**	110.97	**4.80**	16.90	**4.14**	103.54	**2.00**
C60dimer[220]	2.85[g] [BP86-D3(BJ)/def2-TZVP]	**1.91**	**0.69**	**1.91**	**0.69**	**0.90**	**0.79**	**1.13**	**0.95**
L7[221,222]	1.62[h] [B3LYP-NL/def2-TZVP]	3.64	**1.42**	3.64	**1.39**	**1.37**	**1.48**	3.61	**0.82**
S12L[9,11,222]	6.44[h] [BLYP-NL/def2-TZVP]	14.63	**6.41**	14.65	**5.96**	**6.05**	**5.58**	13.51	**6.22**
S30L[219]	6.60[i] [PBE-D3/CBS]	13.07	**5.65**	13.25	**5.23**	**4.80**	**4.74**	11.71	**5.34**
Ni2021[223]	3.20[j,k] [B3LYP-D3(BJ)/triple-ζ]	25.53	5.60	25.53	5.89	6.74	5.98	22.01	10.05

[a)] details about the data sets can be found in Table S2 of the Supporting Information, [b)] from Reference 204, [c)] from Reference 207, [d)] from Reference 209, [e)] from Reference 211, [f)] from Reference 237, [g)] from Reference 220, [h)] from Reference 195, [i)] from Reference 54, [j)] from Reference 223, [k)] aug-cc-pVTZ basis set for six systems and cc-pVTZ basis set for other seven systems, [l)] only non-negatively charged complexes

Two data sets (CE20[207,208], WaterOrg[209]) were used to validate the ACPs for hydrogen bonding interactions. The MAEs of minimal or double-ζ basis set HF-D3 methods for the *Hydrogen-bonding* validation subset ("*Hydrogen-bonding (val)*") are improved on applying the ACPs by 61% (minimal basis set HF-D3) and 84% (double-ζ basis set HF-D3). As in the case of the mixed NCIs, this improvement is close to the one observed in the training set. On the other hand, application of the ACPs to HF-3c neither improves nor deteriorates the MAE for *Hydrogen-bonding (val)*, and the MAE is almost the same as the MAE for hydrogen bonding interactions (0.73 kcal/mol) in the training set. As observed for the training set, ACPs improve methods whose errors are higher, and barely affect methods that already have low MAEs. Comparing to the results obtained with DFT and a large basis set (Table 4), the MAEs of the CE20 data set with ACPs are close to or lower than most of the benchmarked DFT methods with a 6-311+G(3df,2p) basis set in the work of Chan *et al.*[207] Also, the B3LYP-D3(BJ)/6-31+G**-BSIP method that yields results that are close to B3LYP-D3(BJ)/aug-cc-pVQZ has an MAE of 0.44 kcal/mol for WaterOrg data set, which is close that predicted via HF-D3/6-31G*-ACP approach.[209]

The assessment of the ACPs on the validation set also helps us understand what types of interactions are poorly represented in the training set. Based on the analysis performed with the validation set, these interactions are halogen bonding, chalcogen bonding, and close contact repulsions, as discussed below. It should be noted that an assessment of ACPs on interaction types such as π-π stacking, pnicogen bonding, and hydrophobic interactions, discussed earlier for the training set, was not possible in the validation stage because of the scarcity of high-level reference data in the literature.

For the halogen bonding interactions in the validation set ("*Halogen-bonding (val)*" subset), all methods in absence of ACPs show a large over-estimation of the interaction energies. The application of ACPs correct for this over-estimation and lead to a decrease in the MAEs by about 60–76%. Figure 3 shows that the MAEs of the minimal basis set HF-D3 and HF-3c methods without ACPs are almost three times higher than HF-D3/6-31G* (2.01 kcal/mol). Figure 3 also shows that HF-D3/6-31G* has a positive MSE, suggesting an under-estimation in the interaction energies for halogen bonding interactions. This observation for HF-D3/6-31G* is opposite to what was found in the training set. Nonetheless, the spread of errors is decreased when the ACP corrections are used, including HF-D3/6-31G*-ACP, leading to lower SDs than without ACPs.

Model systems representative of chalcogen bonding interactions ("*Chalcogen-bonding*" subset) were absent from the training set. As expected, the improvements in the MAEs when ACPs are applied are not significant for the minimal or double-ζ basis set HF-D3 methods. At the same time, ACPs applied to HF-3c over-estimate the interaction energies and lead to an increase in the MSE and MAE. A slight improvement in the description of chalcogen bonding interactions is observed with ACPs for minimal or double-ζ basis set HF-D3 methods, probably due to the presence of O and S containing complexes in the training set that are not purely chalcogen-bonded. This suggests that increasing the representation of such interactions in the training set could improve the performance and applicability of ACPs. Chalcogen bonding interactions are a difficult test not only for the methods considered in this work but also for many other electronic structure methods. For example, several dispersion-corrected DFT methods tested with ma-def2-QZVPP basis set have MAEs above 1 kcal/mol for the entire CHAL336[211] data set.[211] In this context, Figure 3 suggests that the HF-3c method is the best suited among the minimal basis set HF methods for modeling chalcogen bonding interactions.

Steric repulsive interactions, even though found in some molecules that are forced to be in close contact due to the presence of other attractive interactions or external pressure, seldom occur naturally.[269] Repulsive interactions ("*Repulsive contacts*" subset) are captured well by the minimal or double-ζ basis set HF-D3 and HF-3c methods (MAEs of 0.59–0.89 kcal/mol) and only small improvements are seen with the application of ACPs. Specific subsets for repulsive interactions were missing from our training set. Still, the slight reduction in the MAEs of minimal or double-ζ basis set HF-D3 and HF-3c methods observed with ACPs probably comes from some of the data sets in our training set that contain some data points with repulsive character (e.g. S22x5[135,141,142], S66x8[143–145], S66a8[144], A21x12[3,146,147], and NBC10ext[127,135,148–150]). It should be noted that the MAEs of minimal or double-ζ basis set HF-D3 and HF-3c methods with and without ACPs are lower than the newly reparametrized PM6-D3H4R, DFTB3-D3H4R, PM6-D3H4X, and DFTB3-D3H4X methods (MAEs of 0.94–1.48 kcal/mol). These methods attempt to capture the repulsive interactions via the use of a repulsive energy correction term (parametrized against R160x6[212] and R739x5[213] data sets that constitute the validation *Repulsive contacts* subset) specifically designed for PM6 and DFTB3 methods with the D3H4 correction for dispersion and hydrogen-bonding.[213,270]

Next, we turn our attention to the *$(H_2O)_{20}$ cluster* (H2O20Bind10[215] data set), *C_{60} dimer* (C60dimer[220] data set), and *Large molecule* (L7[221,222], S12L[9,11,222], S30L[219], Ni2021[223] data sets) subsets. These subsets are a good test for ACPs as they contain non-covalently bound complexes that are relatively large and at the same time feature multiple co-operative interactions including one or more of hydrogen bonds, halogen bonds, π-π stacking, H-π, ion-dipole, dispersion, etc. The systems present in *$(H_2O)_{20}$ cluster*, *C_{60} dimer*, and *Large molecule* subsets are known to be challenging not only for minimal or double-ζ basis set HF-D3 and HF-3c methods but also for many other electronic structure methods. The absolute reference interaction energies of many complexes in these subsets range from 25 kcal/mol to 416 kcal/mol. It should also be noted that due to the large size of the systems, the most feasible way to generate the reference data of such systems is via the use of methods like CCSD(T)-F12a (H2O20Bind10), DLPNO-CEPA/1 (C60dimer), and CIM-DLPNO-CCSD(T) (Ni2021) or back-correction of experimental data (*S12L* and *S30L*). The reference data for these data sets are expected to be of lower quality than the others, which are typically calculated at CCSD(T)/CBS. The application of ACPs to minimal or double-ζ basis set HF-D3 and HF-3c methods on *$(H_2O)_{20}$ cluster*, *C_{60} dimer*, and *Large molecule* show a general improvement in the MAEs of the underlying methods. ACPs for minimal basis set HF-D3 reduces the MAEs by about 64–94%, while that for double-ζ basis set HF-D3 reduces the MAEs by about 16–98%. ACPs for HF-3c also reduce the MAEs by about 5–76%. The systems with large errors, higher than ±8 kcal/mol even after the use of ACPs, are, as expected, those that have very large reference energies. Table 4 shows a comparison of MAEs of various HF based approaches with large basis set DFT methods for the H2OBind10, C60dimer, L7, S12L, S30L, and Ni2021 data sets. It can be seen that, with the exception of Ni2021 data set, the MAEs in Table 4 for all other data sets shows that ACPs have a performance similar to large basis set DFT, making the approach particularly promising for modeling interaction energies in large molecular systems.

We now consider the results for the conformational energies evaluated with the subsets "*Small molecule (val)*" and "*Proteinogenic*". The *Small molecule (val)* subset contains conformational energies of various small organic and biaryl drug-like molecules. On the other hand, the *Proteinogenic* subset contains a collection of polypeptide conformers like tripeptides, peptides with disulfide linkages, macrocyclic peptides, and peptide sequences with associated bio-functionality. For the *Small molecule (val)* subset, application of ACPs brings down the MAEs of minimal or double-ζ basis set HF-D3 and HF-3c methods by about 66% (minimal basis set HF-D3), 25% (double-ζ basis set HF-D3), and 50% (HF-3c). For the *Proteinogenic* subset, the improvement in MAEs seen on the application of ACPs is about 40% (minimal basis set HF-D3), 49% (double-ζ basis set HF-D3), and 17% (HF-3c). Figure 3 shows that the spread of errors is more or less symmetric about the zero-error average line, except for some systems with errors higher than ±8 kcal/mol, corresponding to the disulfide linkages which were not present in the training set.

For non-covalent interaction energies of anionic interactions ("*Anionic (val)*" subset), although the application of ACPs leads to a reduction in MAEs of minimal or double-ζ HF-D3 and HF-3c methods by about 24–43%, these MAEs still range between 2.65–5.21 kcal/mol making the overall approach not usable for modeling anionic interactions. The good performance of ACPs for most of the conformational energies in the training and validation sets does not translate to the *Negatively charged (val)* subset. Nevertheless, upon application of ACPs, the MAEs (in kcal/mol) of HF-D3/MINIs is slightly reduced from 4.17 to 3.92 and from 4.74 to 4.03 for HF-3c. The respective MAEs of HF-D3/MINIX-ACP and HF-D3/6-31G*-ACP increase by about 5% and 76%. Similar to the results in the training set, the poor results in the *Anionic (val)* and *Negatively charged (val)* subsets are another indication of the serious problems associated with using minimal and double-ζ basis sets without diffuse functions for negatively charged systems. A possible solution to deal with anionic systems would be to develop ACPs for minimally augmented basis sets[271].

3.3 Performance of ACPs for molecular geometries

One attractive possible use of ACPs is for fast geometry optimizations. To gauge the performance of the various HF based methods for this task, we compared and analyzed the structures obtained after energy relaxation with those obtained using dispersion-corrected DFT methods with large basis sets. The 296 structures used for the test contain both non-covalently bound complexes and single molecules, ranging in size between 2 and 205 atoms. For single molecule structures, we used the equilibrium geometries of small organic molecules taken from our *MOLdef* data set, a variety of organic molecules taken from Reference 272, selected structures from LB12[56] and CLB18[273] data sets, and polypeptide structures from Reference 268. For non-covalently bound structures, we selected the equilibrium complex structures from the A21[147], S66[143], L7[221], and S30L[219] data sets. Wherever high-level geometries were not available, we obtained reference geometries using dispersion-corrected DFT and a reasonably large basis set (CAM-B3LYP-D3(BJ)/6-311++G**) with the "tight" convergence criteria in *Gaussian-16*. All the optimized and reference geometries used for testing are provided in the SI. We used Kabsch's algorithm[274] to compare the optimized structures with the reference.

The results are summarized in Table 5. The table shows that the root-mean-square-deviation (RMSD) of the atomic coordinates for the small basis set HF based methods with ACPs are generally lower than those without ACPs, and this decrease happens for single molecules and non-covalently bound complexes, including charged systems. These results indicate that ACPs are generally able to yield better geometries than the uncorrected methods. The RMSD values for the individual methods and geometries can be found in Table S5 of SI.

To examine the source of the improvement in the molecular geometries upon application of ACPs, we calculated the average error in the intermolecular separation distances for the dimer complexes. We also compared the average error in a few selected bond lengths and angles for the single molecules. Table 5 shows that, on average, the intermolecular separation distances are improved, and the under-estimation in the separation distances yielded by the small basis set HF based methods is corrected by the ACPs, leading to better geometries for non-covalently bound complexes. Furthermore, the average deviations in bond lengths presented in Table 5 indicate that the average deviation of the small basis set HF based methods is between 0.002 Å to 0.092 Å for the selected bonds. It can be seen that the inclusion of ACPs with small basis set HF based methods leads to a better description of bond lengths as the polarity of a bond increases, leading to lower errors in bond lengths (except for HF-3c-ACP). A general improvement in the prediction of bond angles is also observed on application of ACPs. The combination of low average errors in bond lengths and angles leads to better overall geometries of single molecule structures. Despite overall good geometries for both single molecules and complexes, upon application of ACPs some tested geometries tend to have slightly higher RMSDs (greater than 0.7 Å) than the uncorrected methods due to slight deterioration in the bond lengths of C-H and C-C bonds. Such deviations relative to the reference geometry are visible in Table S5 of SI for some purely planar systems and peptides with highly flexible backbones.

Table 5. Results of various methods for equilibrium structures. (RMSD is the root-mean-square deviation in the atomic coordinates, MAE is the mean absolute error, and MSE is the mean signed error.)

	HF-D3/MINIs	**HF-D3/MINIs-ACP**	**HF-D3/MINIX**	**HF-D3/MINIX-ACP**	**HF-3c**	**HF-3c-ACP**	**HF-D3/6-31G***	**HF-D3/6-31G*-ACP**
Overall geometry:								
Mean RMSD (Å) (complexes)	0.326	0.289	0.326	0.284	0.223	0.211	0.320	0.229
Mean RMSD (Å) (single molecules)	0.205	0.167	0.182	0.163	0.158	0.187	0.084	0.049
Mean RMSD (Å) (charged single molecules)	0.759	0.533	0.750	0.513	0.483	0.684	0.516	0.208
Overall mean RMSD (Å)	0.254	0.217	0.240	0.212	0.184	0.196	0.180	0.122
Inter-molecular separation distance[a,b]:								
MAE (Å)	0.222	0.124	0.221	0.126	0.112	0.111	0.157	0.084
MSE (Å)	-0.171	0.015	-0.172	0.017	-0.022	0.002	-0.087	0.005
Selected bond lengths:								
C-H bond (MAE / MSE) (Å)	0.005 / -0.004	0.014 / 0.014	0.005 / -0.004	0.017 / 0.017	0.007 / -0.006	0.018 / 0.018	0.010 / -0.010	0.002 / 0.001
C-C bond (MAE / MSE) (Å)	0.019 / 0.018	0.034 / -0.033	0.019 / 0.018	0.028 / -0.024	0.016 / 0.013	0.028 / -0.023	0.011 / -0.011	0.005 / 0.003
C-N bond (MAE / MSE) (Å)	0.036 / 0.034	0.030 / -0.029	0.036 / 0.034	0.031 / -0.029	0.018 / 0.011	0.033 / -0.030	0.013 / -0.012	0.010 / -0.009
C-O bond (MAE / MSE) (Å)	0.057 / 0.057	0.012 / -0.001	0.057 / 0.057	0.011 / -0.003	0.011 / 0.001	0.014 / -0.008	0.017 / -0.017	0.005 / -0.002
C-F bond (MAE / MSE) (Å)	0.068 / 0.068	0.020 / -0.014	0.068 / 0.068	0.022 / -0.019	0.010 / -0.004	0.030 / -0.030	0.017 / 0.017	0.006 / -0.001
C-Cl bond (MAE / MSE) (Å)	0.092 / 0.092	0.042 / -0.042	0.026 / 0.026	0.039 / -0.039	0.015 / 0.015	0.093 / -0.078	0.017 / -0.017	0.016 / 0.016
Selected bond angles:								
C-C-H angle (MAE / MSE)	0.546 / -0.039	0.365 / 0.070	0.554 / -0.043	0.378 / 0.049	0.421 / -0.063	0.381 / 0.059	0.228 / -0.001	0.197 / 0.007
C-C-C angle (MAE / MSE)	0.792 / -0.434	0.404 / -0.132	0.781 / -0.445	0.443 / -0.216	0.597 / -0.340	0.460 / -0.220	0.336 / -0.193	0.274 / -0.128

	HF-D3/MINIs	**HF-D3/MINIs-ACP**	**HF-D3/MINIX**	**HF-D3/MINIX-ACP**	**HF-3c**	**HF-3c-ACP**	**HF-D3/6-31G***	**HF-D3/6-31G*-ACP**
C-C-N angle (MAE / MSE)	1.385 / -0.444	1.026 / -0.092	1.381 / -0.456	1.114 / -0.350	1.280 / -0.392	1.108 / -0.404	0.583 / -0.171	0.366 / -0.071
C-C-O angle (MAE / MSE)	1.402 / 0.503	0.891 / 0.094	1.325 / 0.491	0.975 / 0.171	1.125 / 0.568	1.108 / 0.248	0.455 / -0.025	0.328 / 0.145
C-C-F angle (MAE / MSE)	0.265 / 0.086	0.218 / 0.045	0.233 / 0.078	0.183 / 0.017	0.267 / 0.154	0.253 / -0.023	0.106 / 0.045	0.224 / 0.133
C-C-Cl angle (MAE / MSE)	0.493 / -0.374	0.377 / 0.078	0.327 / 0.204	0.422 / 0.228	0.435 / 0.323	0.407 / 0.264	0.124 / 0.025	0.305 / -0.121

[a)] calculated as the distance between the centers of mass of each monomer.

[b)] excluding the geometries of the non-dimer complexes from the L7 data set for simplicity.

3.4 Applications of ACPs developed for HF-3c

This section explores the use of HF-3c-ACP for modeling systems where most but not all atoms in the system have an associated ACP. For the atoms for which ACPs are not available, our intention is that HF-3c, which is the overall best of the underlying methods in this work will still give a reasonable description of the system. A particular example of an application where HF-3c-ACP could be used is in modeling metalloproteins where the atoms for which ACPs are unavailable are the metal ion(s) in the active site.

We calculated the interaction energies of two systems from Reference 223 using HF-3c and HF-3c-ACP to demonstrate the above idea. These two systems represent the adsorption of ethanol and benzene with different-sized cluster models of zeolite ZSM-5.[275,276] The ZSM-5 zeolite complexes are mainly composed of H, C, O, and Si atoms for which ACPs are available. However, they also contain an additional aluminum atom. For the ZSM-5 zeolite complexes, the high-level (CIM-DLPNO-CCSD(T)) interaction energy reported in Reference 223 is -12.35 kcal/mol (benzene and zeolite or Benzene-ZSM5) and -36.55 kcal/mol (ethanol and zeolite or Ethanol-ZSM5). The HF-3c approach overestimates the interaction energies and yields -21.38 kcal/mol for Benzene-ZSM5 and -43.86 kcal/mol for Ethanol-ZSM5. The ACPs help reduce the interaction energies over-estimated by HF-3c and brings them closer to the reference: The corrected interaction energies predicted by HF-3c-ACP are -19.41 kcal/mol and -39.75 kcal/mol, respectively.

We now explore the same idea by taking some subsets of the validation set and purposefully applying only part of the available ACPs so that not all atoms in the system have an associated correction. Table 6 presents a summary of the MAEs using various methods for two data sets from the validation set: the DES15K[205] set of non-covalent interaction energies (11,474 data points) and the Torsion30[227] set of molecular conformational energies (2,107 data points). The table shows that the MAEs of the HF/MINIX method are 4.83 and 0.92 kcal/mol for the DES15K and Torsion30 data sets, respectively. Using the HF-3c method, the MAE decreases for DES15K (2.14 kcal/mol) and increases to 1.18 kcal/mol for Torsion30. Table 6 shows that using HF-3c-ACP but applying ACPs only to hydrogen and one of the non-hydrogen atoms indicates that ACPs improve the results progressively as the correction is applied to more atoms in the system. If only hydrogen and one of the non-hydrogen atoms are corrected, the performance of HF-3c-ACP is similar to HF-3c. If the correction is applied to hydrogen and two non-hydrogen atoms, most atoms are corrected and the MAEs decrease substantially and resemble HF-3c-ACP, in which all atoms receive an ACP. Therefore, we conclude that the application of ACPs is, in general, beneficial, and greater performance is obtained as more atoms receive an ACP, so the use of ACPs is recommended even in systems containing atoms for which ACPs are not available.

Table 6. Mean absolute error (MAE) for the DES15K data set of non-covalent interaction energies and Torsion30 of conformational energies for HF/MINIX, HF-3c, and HF-3c with application of ACPs to various atoms.

Subset	**Mean absolute error (in kcal/mol)**									
	HF/MINIX	**HF-3c**	**HF-3c with H & O ACPs**	**HF-3c with H & N ACPs**	**HF-3c with H & C ACPs**	**HF-3c with H, N, & O ACPs**	**HF-3c with H, C, & O ACPs**	**HF-3c with H, C, & N ACPs**	**HF-3c with H, C, N, & O ACPs**	**HF-3c-ACP**
DES15K[a]	4.83	2.14	1.95	2.24	2.17	1.83	1.78	2.07	1.68	1.32
Torsion30[b]	0.92	1.18	1.20	1.09	0.70	1.06	0.70	0.58	0.57	0.59

[a)] atom frequency: H = 138132, C = 62376, O = 11604, N = 9250, other (F, P, S, Cl) = 7936.

[b)] atom frequency: C = 23744, H = 18473, N = 4143, O = 985, other (S) = 70.

4. Summary and Outlook

An important field of research in modern computational chemistry is the development of new quantum mechanical methods that are accurate and can be applied to model large molecular systems. Small basis set Hartree–Fock (HF) methods are orders of magnitude less expensive than more accurate nearly complete basis set wavefunction theory or DFT methods, but suffer from basis set incompleteness error and lack of electronic correlation. Provided these shortcomings can be addressed, such methods could be applied for modeling large molecular systems as well as for routine applications of fast geometry optimizations, conformational exploration, and prediction of non-covalent interaction strengths.

In this work, we show that HF with small and minimal basis sets can be effectively corrected by applying atom-centered potentials (ACPs, one-electron potentials similar to effective-core potentials) that are designed to correct for the inaccuracies in the underlying method. Four new sets of ACPs were developed for use with HF-D3 and small basis sets (MINIs, MINIX, 6-31G*) and HF-3c. The advantages of ACPs include that they can be used in most computational chemistry software packages without changes to the code and that they incur only a modest computational cost. The ACPs developed in this work apply to ten elements (H, B, C, N, O, F, Si, P, S, Cl), and our purpose is that the presented ACPs serve to address problems in organic chemistry and biochemistry. For the occasional system containing atoms for which no ACPs are available, we have shown that the improvement of the performance of the underlying method is progressive with the number of atoms where ACPs have been applied. Therefore, the use of ACPs is beneficial even if some atoms are not corrected, and in this case, we recommend the use of HF-3c-ACP. We anticipate that the ACP based approaches developed in this work will allow efficient and accurate modeling of biomolecules such as proteins, nucleic acids, carbohydrates, lipids, and other molecules containing B, Si, and halogen atoms such as covalent organic frameworks, functionalized polyaromatic hydrocarbons, functionalized silicon surfaces, and more.

The ACPs were developed by using a large training set of 73,832 data points calculated at a very high level of theory (CCSD(T)/CBS, in general). The training set contains a mixture of non-covalent interaction energies, molecular conformational energies, and molecular deformation energies. We expected that the size of the training set ensures the robustness and applicability of the ACPs. To test this, we validated the new ACPs on a validation set with 32,047 data points. The assessment of minimal and double-ζ basis set HF-D3 and HF-3c methods, before and after the application of their corresponding ACPs showed that ACPs lower the MAEs of most subsets in the training set and that this good performance is carried over to the validation set with approximately the same performance in terms of average error. Relative to the uncorrected methods, ACP-corrected approaches improve the prediction of non-covalent interaction energies and molecular conformational energies. Furthermore, the addition of molecular deformation energies to the training set results in an improvement of the equilibrium molecular structures upon application of the ACPs. However, ACP-corrected methods showed relatively poor performance for some interaction types that were not part of the training set, such as chalcogen bonding or repulsive contacts, indicating that more diverse systems need to be included in the training set for greater robustness of the resulting methods.

Our analysis of representative data sets indicates that our ACP-corrected methods yield results similar to almost complete basis set DFT methods, naturally at a much lower computational cost. Nonetheless, there remains a limitation regarding the description of negatively charged systems probably caused by the lack of diffuse functions in the basis sets employed. In spite of this, our ACPs offer a modest improvement even in this case. ACPs for small basis sets that include some diffuse functions are currently under development. We are also currently working on expanding the set of ACPs to DFT-D3 methods with small basis sets for prediction of accurate thermochemical properties along with non-covalent properties. Despite the limitation, we have shown that ACPs provide a way of developing methods that combine low-cost with robustness and wide applicability. We anticipate that ACPs will be useful to practitioners interested in modeling large systems or other time-intensive applications.

References

(1) E. G. Hohenstein and C. D. Sherrill, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**, 304–326.

(2) S. M. Bachrach, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2014, **4**, 482–487.

(3) J. Řezáč and P. Hobza, *J. Chem. Theory Comput.*, 2013, **9**, 2151–2155.

(4) Y. S. Al-Hamdani and A. Tkatchenko, *J. Chem. Phys.*, 2019, **150**, 10901.

(5) T. Helgaker, P. Jørgensen and J. Olsen, in *Molecular Electronic-Structure Theory*, John Wiley & Sons, Ltd, Chichester, UK, 2014, pp. 817–883.

(6) Q. Cui, *J. Chem. Phys.*, 2016, **145**, 140901.

(7) L. E. Ratcliff, S. Mohr, G. Huhs, T. Deutsch, M. Masella and L. Genovese, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2017, **7**, e1290.

(8) K. E. Riley and P. Hobza, *Wiley Interdiscip. Rev.: Comput Mol Sci*, 2011, **1**, 3–17.

(9) A. Ambrosetti, D. Alfè, R. A. DiStasio Jr. and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2014, **5**, 849–855.

(10) A. Otero-de-la-Roza and E. R. Johnson, *J. Chem. Theory Comput.*, 2015, **11**, 4033–4040.

(11) T. Risthaus and S. Grimme, *J. Chem. Theory Comput.*, 2013, **9**, 1580–1591.

(12) J. G. Brandenburg, K. Burke, B. Civalleri, D. J. Cole, G. Csányi, G. David, N. I. Gidopoulos, D. Gowland, T. Helgaker, M. F. Herbst, B. Hourahine, T. J. P. Irons, C. R. Jacob, P. F. Loos, N. Mehta, M. R. Mulay, J. Neugebauer, K. Pernal, A. Pribram-Jones, P. Romaniello, M. R. Ryder, A. Savin, D. Sirbu, C. K. Skylaris, D. G. Truhlar, J. Wetherell and W. Yang, *Faraday Discuss.*, 2020, **224**, 309–332.

(13) Y. S. Al-Hamdani, P. R. Nagy, A. Zen, D. Barton, M. Kállay, J. G. Brandenburg and A. Tkatchenko, *Nat. Comm.*, 2021, **12**, 3927.

(14) J. Antony, R. Sure and S. Grimme, *Chem. Commun.*, 2015, **51**, 1764–1774.

(15) C. D. Sherrill, *J. Chem. Phys.*, 2010, **132**, 110902.

(16) T. S. Hofer, *Front. Chem.*, 2013, **1**, 6.

(17) S. Grimme and P. R. Schreiner, *Angew. Chemie Int. Ed.*, 2018, **57**, 4170–4176.

(18) K. N. Houk and F. Liu, *Acc. Chem. Res.*, 2017, **50**, 539–543.

(19) K. M. Merz, *Acc. Chem. Res.*, 2014, **47**, 2804–2811.

(20) M. S. Gordon, D. G. Fedorov, S. R. Pruitt and L. V. Slipchenko, *Chem. Rev.*, 2012, **112**, 632–672.

(21) M. A. Collins and R. P. A. Bettens, *Chem. Rev.*, 2015, **115**, 5607–5642.

(22) S. Li, W. Li and J. Ma, *Acc. Chem. Res.*, 2014, **47**, 2712–2720.

(23) W. Li, H. Dong, J. Ma and S. Li, *Acc. Chem. Res.*, 2021, **54**, 169–181.

(24) K. Raghavachari and A. Saha, *Chem. Rev.*, 2015, **115**, 5643–5677.

(25) X. He, T. Zhu, X. Wang, J. Liu and J. Z. H. Zhang, *Acc. Chem. Res.*, 2014, **47**, 2748–2757.

(26) R. O. Ramabhadran and K. Raghavachari, *Acc. Chem. Res.*, 2014, **47**, 3596–3604.

(27) M. A. Collins, M. W. Cvitkovic and R. P. A. Bettens, *Acc. Chem. Res.*, 2014, **47**, 2776–2785.

(28) S. R. Pruitt, C. Bertoni, K. R. Brorsen and M. S. Gordon, *Acc. Chem. Res.*, 2014, **47**, 2786–2794.

(29) R. Sure, J. G. Brandenburg and S. Grimme, *ChemistryOpen*, 2016, **5**, 94–109.

(30) L. Goerigk, C. A. Collyer and J. R. Reimers, *J. Phys. Chem. B*, 2014, **118**, 14612–14626.

(31) A. S. Christensen, T. Kubař, Q. Cui and M. Elstner, *Chem. Rev.*, 2016, **116**, 5301–5337.

(32) N. D. Yilmazer and M. Korth, *Comput. Struct. Biotechnol. J.*, 2015, **13**, 169–175.

(33) P. O. Dral, X. Wu, L. Spörkel, A. Koslowski, W. Weber, R. Steiger, M. Scholten and W. Thiel, *J. Chem. Theory Comput.*,

2016, **12**, 1082–1096.

(34) W. Thiel, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2014, **4**, 145–157.

(35) K. T. Throssell, PhD thesis, Wesleyan University, 2018.

(36) J. J. P. Stewart, *J. Mol. Model.*, 2013, **19**, 1–32.

(37) R. Sure and S. Grimme, *J. Comput. Chem.*, 2013, **34**, 1672–1685.

(38) S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.

(39) H. Kruse and S. Grimme, *J. Chem. Phys.*, 2012, **136**, 154101.

(40) S. Grimme, C. Bannwarth and P. Shushkov, *J. Chem. Theory Comput.*, 2017, **13**, 1989–2009.

(41) C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.

(42) S. Spicher and S. Grimme, *Angew. Chemie Int. Ed.*, 2020, **59**, 15665–15673.

(43) C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher and S. Grimme, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2021, **11**, e1493.

(44) P. Pracht, E. Caldeweyher, S. Ehlert and S. Grimme, 2019, *ChemRxiv*:10.26434/chemrxiv.8326202.v1.

(45) T. Z. H. Gani and H. J. Kulik, *J. Chem. Theory Comput.*, 2016, **12**, 5931–5945.

(46) R. Alizadegan, K. J. Hsia and T. J. Martinez, *J. Chem. Phys.*, 2010, **132**, 034101.

(47) J. Garcia and K. Szalewicz, *J. Phys. Chem. A*, 2020, **124**, 1196–1203.

(48) Y. Chen, L. Zhang, H. Wang and W. Weinan, *J. Phys. Chem. A*, 2020, **124**, 7155–7165.

(49) A. Altun, F. Neese and G. Bistoni, *J. Chem. Theory Comput.*, 2019, **15**, 5894–5907.

(50) G. M. J. Barca, J. L. Galvez-Vallejo, D. L. Poole, A. P. Rendell and M. S. Gordon, *J. Chem. Theory Comput.*, 2020, **16**, 7232–7238.

(51) J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.

(52) J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev and A. E. Roitberg, *Nat. Commun.*, 2019, **10**, 1–8.

(53) C. Devereux, J. S. Smith, K. K. Davis, K. Barros, R. Zubatyuk, O. Isayev and A. E. Roitberg, *J. Chem. Theory Comput.*, 2020, **16**, 4192–4202.

(54) J. G. Brandenburg, M. Hochheim, T. Bredow and S. Grimme, *J. Phys. Chem. Lett.*, 2014, **5**, 4275–4284.

(55) M. A. Iron and T. Janes, *J. Phys. Chem. A*, 2019, **123**, 3761–3781.

(56) S. Grimme, J. G. Brandenburg, C. Bannwarth and A. Hansen, *J. Chem. Phys.*, 2015, **143**, 054107.

(57) J. G. Brandenburg, C. Bannwarth, A. Hansen and S. Grimme, *J. Chem. Phys.*, 2018, **148**, 064104.

(58) S. Grimme, A. Hansen, S. Ehlert and J.-M. Mewes, *J. Chem. Phys.*, 2021, **154**, 064103.

(59) P. Pracht, D. F. Grant and S. Grimme, *J. Chem. Theory Comput.*, 2020, **16**, 7044–7060.

(60) J. G. Brandenburg, E. Caldeweyher and S. Grimme, *Phys. Chem. Chem. Phys.*, 2016, **18**, 15519–15523.

(61) E. Caldeweyher and J. G. Brandenburg, *J. Phys. Condens. Matter*, 2018, **30**, 213001.

(62) G. A. DiLabio, in *Non-covalent Interactions in Quantum Chemistry and Physics: Theory and Applications*, ed. A. Otero-de-la-Roza and G. A. DiLabio, Elsevier Inc., 2017, pp. 221–240.

(63) V. K. Prasad, A. Otero-de-la-Roza and G. A. DiLabio, *J. Chem. Theory Comput.*, 2018, **14**, 726–738.

(64) A. Otero-De-La-Roza and G. A. Dilabio, *J. Chem. Theory Comput.*, 2020, **16**, 4176–4191.

(65) A. Otero-de-la-Roza and G. A. DiLabio, *J. Chem. Theory Comput.*, 2017, **13**, 3505–3524.

(66) J. A. van Santen and G. A. DiLabio, *J. Phys. Chem. A*, 2015, **119**, 6703–6713.

(67) G. A. DiLabio and M. Koleini, *J. Chem. Phys.*, 2014, **140**, 18A542.

(68) G. A. DiLabio, M. Koleini and E. Torres, *Theor. Chem. Acc.*, 2013, **132**, 1389.

(69) E. Torres and G. A. DiLabio, *J. Phys. Chem. Lett.*, 2012, **3**, 1738–1744.

(70) I. D. Mackie and G. A. DiLabio, *J. Phys. Chem. A*, 2008, **112**, 10968–10976.

(71) G. A. DiLabio, *Chem. Phys. Lett.*, 2008, **455**, 348–353.

(72) I. D. Mackie and G. A. DiLabio, *Phys. Chem. Chem. Phys.*, 2010, **12**, 6092.

(73) E. Torres and G. A. DiLabio, *J. Chem. Theory Comput.*, 2013, **9**, 3342–3349.

(74) I. D. Mackie and G. A. DiLabio, *Phys. Chem. Chem. Phys.*, 2011, **13**, 2780–2787.

(75) J. D. Holmes, A. Otero-de-la-Roza and G. A. DiLabio, *J. Chem. Theory Comput.*, 2017, **13**, 4205–4215.

(76) X. Cao and M. Dolg, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**, 200–210.

(77) M. Dolg and X. Cao, *Chem. Rev.*, 2012, **112**, 403–480.

(78) R. Tibshirani, *J. R. Stat. Soc. Ser. B (Statistical Methodology)*, 2011, **73**, 273–282.

(79) R. Tibshirani, *J. R. Stat. Soc. Ser. B*, 1996, **58**, 267–288.

(80) M. R. Osborne, B. Presnell and B. A. Turlach, *J. Comput. Graph. Stat.*, 2000, **9**, 319–337.

(81) H. Tatewaki and S. Huzinaga, *J. Comput. Chem.*, 1980, **1**, 205–228.

(82) M. M. Francl, W. J. Pietro, W. J. Hehre, J. S. Binkley, M. S. Gordon, D. J. DeFrees and J. A. Pople, *J. Chem. Phys.*, 1982, **77**, 3654–3665.

(83) P. C. Hariharan and J. A. Pople, *Theor. Chim. Acta*, 1973, **28**, 213–222.

(84) S. Grimme, S. Ehrlich and L. Goerigk, *J. Comput. Chem.*, 2011, **32**, 1456–1465.

(85) E. R. Johnson and A. D. Becke, *J. Chem. Phys.*, 2006, **124**, 174104.

(86) M. Habgood, T. James and A. Heifetz, in *Methods in Molecular Biology*, Humana Press Inc., 2020, vol. 2114, pp. 207–229.

(87) P. Pracht, F. Bohle and S. Grimme, *Phys. Chem. Chem. Phys.*, 2020, **22**, 7169–7192.

(88) P. C. D. Hawkins, *J. Chem. Inf. Model.*, 2017, **57**, 1747–1756.

(89) S. Wang, J. Witek, G. A. Landrum and S. Riniker, *J. Chem. Inf. Model.*, 2020, **60**, 2044–2058.

(90) T. Rezai, J. E. Bock, M. V. Zhou, C. Kalyanaraman, R. S. Lokey and M. P. Jacobson, *J. Am. Chem. Soc.*, 2006, **128**, 14073–14080.

(91) V. Poongavanam, E. Danelius, S. Peintner, L. Alcaraz, G. Caron, M. D. Cummings, S. Wlodek, M. Erdelyi, P. C. D. Hawkins, G. Ermondi and J. Kihlberg, *ACS Omega*, 2018, **3**, 11742–11757.

(92) D. B. Diaz, S. D. Appavoo, A. F. Bogdanchikova, Y. Lebedev, T. J. McTiernan, G. dos Passos Gomes and A. K. Yudin, *Nat. Chem.*, 2021, **13**, 218–225.

(93) I. Kolossváry and W. C. Guida, *J. Am. Chem. Soc.*, 1996, **118**, 5011–5019.

(94) O. Gutten, D. Bím, J. Řezáč and L. Rulíšek, *J. Chem. Inf. Model.*, 2018, **58**, 48–60.

(95) I. Saha, E. K. Dang, D. Svatunek, K. N. Houk and P. G. Harran, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 24679–24690.

(96) K. T. Butler, F. J. Luque and X. Barril, *J. Comput. Chem.*, 2009, **30**, 601–610.

(97) I. Sorokina and A. Mushegian, *Biol. Direct*, 2016, **11**, 1–5.

(98) M. Culka and L. Rulíšek, *J. Phys. Chem. B*, 2019, **123**, 6453–6461.

(99) M. Culka, J. Galgonek, J. Vymětal, J. Vondrášek and L. Rulíšek, *J. Phys. Chem. B*, 2019, **123**, 1215–1227.

(100) C.-A. Mattelaer, H.-P. Mattelaer, J. Rihon, M. Froeyen and E. Lescrinier, *J. Chem. Theory Comput.*, 2021, **17**, 3814–3823.

(101) M. Huang, T. J. Giese, T. S. Lee and D. M. York, *J. Chem. Theory Comput.*, 2014, **10**, 1538–1545.

(102) A. Di Fenza, A. Heine, U. Koert and G. Klebe, *ChemMedChem*, 2007, **2**, 297–308.

(103) G. B. McGaughey, M. Gagné and A. K. Rappé, *J. Biol. Chem.*, 1998, **273**, 15458–15463.

(104) N. Sal-Man, D. Gerber, I. Bloch and Y. Shai, *J. Biol. Chem.*, 2007, **282**, 19753–19761.

(105) M. K. Ravva, C. Risko and J. L. Brédas, in *Non-Covalent Interactions in Quantum Chemistry and Physics: Theory and Applications*, Elsevier Inc., 2017, pp. 277–302.

(106) J. Černý, M. Kabeláč and P. Hobza, *J. Am. Chem. Soc.*, 2008, **130**, 16055–16059.

(107) H. Karabiyik, R. Sevinçek and H. Karabiyik, *Phys. Chem. Chem. Phys.*, 2014, **16**, 15527–15538.

(108) K. A. Wilson and S. D. Wetmore, in *Challenges and Advances in Computational Chemistry and Physics*, Springer, 2015, vol. 19, pp. 501–532.

(109) L. M. Salonen, M. Ellermann and F. Diederich, *Angew. Chemie - Int. Ed.*, 2011, **50**, 4808–4842.

(110) H. J. Schneider, *Angew. Chemie Int. Ed.*, 2009, **48**, 3924–3977.

(111) J. H. Deng, J. Luo, Y. L. Mao, S. Lai, Y. N. Gong, D. C. Zhong and T. B. Lu, *Sci. Adv.*, 2020, **6**, eaax9976.

(112) J. wun Hwang, P. Li and K. D. Shimizu, *Org. Biomol. Chem.*, 2017, **15**, 1554–1564.

(113) K. Berka, R. A. Laskowski, P. Hobza and J. Vondrášek, *J. Chem. Theory Comput.*, 2010, **6**, 2191–2203.

(114) J. L. MacCallum, W. F. Drew Bennett and D. Peter Tieleman, *Biophys. J.*, 2008, **94**, 3393–3404.

(115) M. N. Mbaye, Q. Hou, S. Basu, F. Teheux, F. Pucci and M. Rooman, *Sci. Rep.*, 2019, **9**, 1–14.

(116) E. Busseron, Y. Ruff, E. Moulin and N. Giuseppone, *Nanoscale*, 2013, **5**, 7098–7140.

(117) G. R. Desiraju and T. Steiner, in *The Weak Hydrogen Bond In Strcutural Chemistry and Biology*, Oxford University Press, Oxford and New York, 1999.

(118) D. Herschlag and M. M. Pinney, *Biochemistry*, 2018, **57**, 3338–3352.

(119) A. Bauzá, P. M. Deyà and A. Frontera, in *Challenges and Advances in Computational Chemistry and Physics*, Springer, 2015, vol. 19, pp. 471–500.

(120) B. L. Schottel, H. T. Chifotides and K. R. Dunbar, *Chem. Soc. Rev.*, 2008, **37**, 68–83.

(121) X. Lucas, A. Bauzá, A. Frontera and D. Quiñonero, *Chem. Sci.*, 2016, **7**, 1038–1050.

(122) S. Z. Borozan, M. V. Zlatović and S. Stojanović, *J. Biol. Inorg. Chem.*, 2016, **21**, 357–368.

(123) J. M. Sanders, *J. Phys. Chem. A*, 2010, **114**, 9205–9211.

(124) R. M. Parrish and C. D. Sherrill, *J. Am. Chem. Soc.*, 2014, **136**, 17386–17389.

(125) S. N. Steinmann and C. Corminboeuf, *J. Chem. Theory Comput.*, 2012, **8**, 4305–4316.

(126) D. L. Crittenden, *J. Phys. Chem. A*, 2009, **113**, 1663–1669.

(127) D. G. A. Smith, L. A. Burns, K. Patkowski and C. D. Sherrill, *J. Phys. Chem. Lett.*, 2016, **7**, 2197–2203.

(128) L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi and S. Grimme, *Phys. Chem. Chem. Phys.*, 2017, **19**, 32184–32215.

(129) S. Tsuzuki, K. Honda, T. Uchimaru and M. Mikami, *J. Chem. Phys.*, 2006, **124**, 114304.

(130) J. Granatier, M. Pitoňák and P. Hobza, *J. Chem. Theory Comput.*, 2012, **8**, 2282–2292.

(131) D. Setiawan, E. Kraka and D. Cremer, *J. Phys. Chem. A*, 2015, **119**, 1642–1656.

(132) J. G. Hill and A. C. Legon, *Phys. Chem. Chem. Phys.*, 2015, **17**, 858–867.

(133) J. Řezáč, K. E. Riley and P. Hobza, *J. Chem. Theory Comput.*, 2012, **8**, 4285–4292.

(134) K. S. Thanthiriwatte, E. G. Hohenstein, L. A. Burns and C. D. Sherrill, *J. Chem. Theory Comput.*, 2011, **7**, 88–96.

(135) M. S. Marshall, L. A. Burns and C. D. Sherrill, *J. Chem. Phys.*, 2011, **135**, 194102.

(136) J. Řezáč, J. Fanfrlík, D. Salahub and P. Hobza, *J. Chem. Theory Comput.*, 2009, **5**, 1749–1760.

(137) V. M. Miriyala and J. Řezáč, *J. Comput. Chem.*, 2017, **38**, 688–697.

(138) J. Řezáč and P. Hobza, *J. Chem. Theory Comput.*, 2012, **8**, 141–151.

(139) J. Řezáč, *J. Chem. Theory Comput.*, 2020, **16**, 2355–2368.

(140) J. Řezáč, *J. Chem. Theory Comput.*, 2020, **16**, 6305–6316.

(141) P. Jurečka, J. Šponer, J. Černý and P. Hobza, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1985–1993.

(142) L. Gráfová, M. Pitoňák, J. Řezáč and P. Hobza, *J. Chem. Theory Comput.*, 2010, **6**, 2365–2376.

(143) J. Řezáč, K. E. Riley and P. Hobza, *J. Chem. Theory Comput.*, 2011, **7**, 2427–2438.

(144) J. Řezáč, K. E. Riley and P. Hobza, *J. Chem. Theory Comput.*, 2011, **7**, 3466–3470.

(145) G. A. Dilabio, E. R. Johnson and A. Otero-de-la-Roza, *Phys. Chem. Chem. Phys.*, 2013, **15**, 12821–12828.

(146) J. Řezáč, M. Dubecký, P. Jurečka and P. Hobza, *Phys. Chem. Chem. Phys.*, 2015, **17**, 19268–19277.

(147) J. Witte, M. Goldey, J. B. Neaton and M. Head-Gordon, *J. Chem. Theory Comput.*, 2015, **11**, 1481–1492.

(148) C. David Sherrill, T. Takatani and E. G. Hohenstein, *J. Phys. Chem. A*, 2009, **113**, 10146–10159.

(149) E. G. Hohenstein and C. D. Sherrill, *J. Phys. Chem. A*, 2009, **113**, 878–886.

(150) T. Takatani and C. David Sherrill, *Phys. Chem. Chem. Phys.*, 2007, **9**, 6106–6114.

(151) J. Řezáč, Y. Huang, P. Hobza and G. J. O. Beran, *J. Chem. Theory Comput.*, 2015, **11**, 3065–3079.

(152) K. L. Copeland and G. S. Tschumper, *J. Chem. Theory Comput.*, 2012, **8**, 1646–1656.

(153) L. A. Burns, J. C. Faver, Z. Zheng, M. S. Marshall, D. G. A. Smith, K. Vanommeslaeghe, A. D. MacKerell, K. M. Merz and C. D. Sherrill, *J. Chem. Phys.*, 2017, **147**, 161727.

(154) J. Černý, B. Schneider and L. Biedermannová, *Phys. Chem. Chem. Phys.*, 2017, **19**, 17094–17102.

(155) J. C. Faver, M. L. Benson, X. He, B. P. Roberts, B. Wang, M. S. Marshall, M. R. Kennedy, C. D. Sherrill and K. M. Merz, *J. Chem. Theory Comput.*, 2011, **7**, 790–797.

(156) K. Kříž and J. Řezáč, *J. Chem. Inf. Model.*, 2020, **60**, 1453–1460.

(157) K. U. Lao, R. Schäffer, G. Jansen and J. M. Herbert, *J. Chem. Theory Comput.*, 2015, **11**, 2473–2486.

(158) D. Jakubec, J. Hostaš, R. A. Laskowski, P. Hobza and J. Vondrášek, *J. Chem. Theory Comput.*, 2015, **11**, 1939–1948.

(159) J. Hostaš, D. Jakubec, R. A. Laskowski, R. Gnanasekaran, J. Řezáč, J. Vondrášek and P. Hobza, *J. Chem. Theory Comput.*, 2015, **11**, 4086–4092.

(160) D. Jakubec, R. A. Laskowski and J. Vondrasek, *PLoS One*, 2016, **11**, e0158704.

(161) O. A. Stasyuk, D. Jakubec, J. Vondrášek and P. Hobza, *J. Chem. Theory Comput.*, 2017, **13**, 877–885.

(162) S. Kozmon, R. Matuška, V. Spiwok and J. Koča, *Chem. - A Eur. J.*, 2011, **17**, 5680–5690.

(163) S. Kozmon, R. Matuška, V. Spiwok and J. Koča, *Phys. Chem. Chem. Phys.*, 2011, **13**, 14215–14222.

(164) I. M. Stanković, J. P. Blagojević Filipović and S. D. Zarić, *Int. J. Biol. Macromol.*, 2020, **157**, 1–9.

(165) M. Kumari, R. B. Sunoj and P. V. Balaji, *Org. Biomol. Chem.*, 2012, **10**, 4186–4200.

(166) H. Kruse, P. Banáš and J. Šponer, *J. Chem. Theory Comput.*, 2019, **15**, 95–115.

(167) T. M. Parker and C. D. Sherrill, *J. Chem. Theory Comput.*, 2015, **11**, 4197–4204.

(168) P. Banáš, A. Mládek, M. Otyepka, M. Zgarbová, P. Jurečka, D. Svozil, F. Lankaš and J. Šponer, *J. Chem. Theory Comput.*, 2012, **8**, 2448–2460.

(169) M. Kabeláč, H. Valdes, E. C. Sherer, C. J. Cramer and P. Hobza, *Phys. Chem. Chem. Phys.*, 2007, **9**, 5000–5008.

(170) D. G. A. Smith and K. Patkowski, *J. Phys. Chem. C*, 2014, **118**, 544–550.

(171) D. G. A. Smith and K. Patkowski, *J. Chem. Theory Comput.*, 2013, **9**, 370–389.

(172) K. D. Vogiatzis, W. Klopper and J. Friedrich, *J. Chem. Theory Comput.*, 2015, **11**, 1574–1584.

(173) D. G. A. Smith and K. Patkowski, *J. Phys. Chem. C*, 2015, **119**, 4934–4948.

(174) S. Li, D. G. A. Smith and K. Patkowski, *Phys. Chem. Chem. Phys.*, 2015, **17**, 16560–16574.

(175) W. Li, S. Grimme, H. Krieg, J. Möllmann and J. Zhang, *J. Phys. Chem. C*, 2012, **116**, 8865–8871.

(176) B. Temelso, K. A. Archer and G. C. Shields, *J. Phys. Chem. A*, 2011, **115**, 12034–12046.

(177) E. M. Mas, R. Bukowski, K. Szalewicz, G. C. Groenenboom, P. E. S. Wormer and A. Van Der Avoird, *J. Chem. Phys.*, 2000, **113**, 6687–6701.

(178) R. Bukowski, K. Szalewicz, G. C. Groenenboom and A. Van Der Avoird, *Science*, 2007, **315**, 1249–1252.

(179) R. Bukowski, K. Szalewicz, G. C. Groenenboom and A. Van Der Avoird, *J. Chem. Phys.*, 2008, **128**, 094313.

(180) B. J. Mintz and J. M. Parks, *J. Phys. Chem. A*, 2012, **116**, 1086–1092.

(181) D. I. Sharapa, A. Genaev, L. Cavallo and Y. Minenkov, *ChemPhysChem*, 2018, **20**, 92–102.

(182) B. D. Sellers, N. C. James and A. Gobbi, *J. Chem. Inf. Model.*, 2017, **57**, 1265–1275.

(183) U. R. Fogueri, S. Kozuch, A. Karton and J. M. L. Martin, *J. Phys. Chem. A*, 2013, **117**, 2269–2277.

(184) D. N. Tahchieva, D. Bakowies, R. Ramakrishnan and O. A. Von Lilienfeld, *J. Chem. Theory Comput.*, 2018, **14**, 4806–4817.

(185) D. Folmsbee and G. Hutchison, *Int. J. Quantum Chem.*, 2021, **121**, e26381.

(186) J. S. Smith, R. Zubatyuk, B. Nebgen, N. Lubbers, K. Barros, A. E. Roitberg, O. Isayev and S. Tretiak, *Sci. Data*, 2020, **7**, 1–10.

(187) V. K. Prasad, A. Otero-de-la-Roza and G. A. DiLabio, *Sci. Data*, 2019, **6**, 180310.

(188) L. Goerigk, A. Karton, J. M. L. Martin and L. Radom, *Phys. Chem. Chem. Phys.*, 2013, **15**, 7028.

(189) H. Valdes, K. Pluháčková, M. Pitonák, J. Řezáč and P. Hobza, *Phys. Chem. Chem. Phys.*, 2008, **10**, 2747.

(190) M. K. Kesharwani, A. Karton and J. M. L. Martin, *J. Chem. Theory Comput.*, 2016, **12**, 444–454.

(191) A. Mládek, M. Krepl, D. Svozil, P. Čech, M. Otyepka, P. Banáš, M. Zgarbová, P. Jurečka and J. Šponer, *Phys. Chem. Chem. Phys.*, 2013, **15**, 7295–7310.

(192) A. Mládek, P. Banáš, P. Jurečka, M. Otyepka, M. Zgarbová and J. Šponer, *J. Chem. Theory Comput.*, 2014, **10**, 463–480.

(193) H. Kruse, A. Mladek, K. Gkionis, A. Hansen, S. Grimme and J. Sponer, *J. Chem. Theory Comput.*, 2015, **11**, 4972–4991.

(194) G. I. Csonka, A. D. French, G. P. Johnson and C. A. Stortz, *J. Chem. Theory Comput.*, 2009, **5**, 679–692.

(195) B. Chan, *J. Phys. Chem. A*, 2020, **124**, 582–590.

(196) W. M. C. Sameera and D. A. Pantazis, *J. Chem. Theory Comput.*, 2012, **8**, 2630–2645.

(197) M. Marianski, A. Supady, T. Ingram, M. Schneider and C. Baldauf, *J. Chem. Theory Comput.*, 2016, **12**, 6157–6168.

(198) D. Gruzman, A. Karton and J. M. L. Martin, *J. Phys. Chem. A*, 2009, **113**, 11974–11983.

(199) S. Kozuch, S. M. Bachrach and J. M. L. Martin, *J. Phys. Chem. A*, 2014, **118**, 293–303.

(200) J. M. L. Martin, *J. Phys. Chem. A*, 2013, **117**, 3118–3132.

(201) B. Temelso, K. L. Klein, J. W. Mabey, C. Pérez, B. H. Pate, Z. Kisiel and G. C. Shields, *J. Chem. Theory Comput.*, 2018, **14**, 1141–1153.

(202) B. J. Smith, D. J. Swanton, J. A. Pople, H. F. Schaefer and L. Radom, *J. Chem. Phys.*, 1990, **92**, 1240–1247.

(203) G. S. Tschumper, M. L. Leininger, B. C. Hoffman, E. F. Valeev, H. F. Schaefer and M. Quack, *J. Chem. Phys.*, 2002, **116**, 690–701.

(204) D. E. Taylor, J. G. Ángyán, G. Galli, C. Zhang, F. Gygi, K. Hirao, J. W. Song, K. Rahul, O. Anatole Von Lilienfeld, R. Podeszwa, I. W. Bulik, T. M. Henderson, G. E. Scuseria, J. Toulouse, R. Peverati, D. G. Truhlar and K. Szalewicz, *J. Chem. Phys.*, 2016, **145**, 124105.

(205) A. G. Donchev, A. G. Taube, E. Decolvenaere, C. Hargus, R. T. McGibbon, K.-H. Law, B. A. Gregersen, J.-L. Li, K. Palmo, K. Siva, M. Bergdorf, J. L. Klepeis and D. E. Shaw, *Sci. Data*, 2021, **8**, 1–9.

(206) Z. M. Sparrow, B. G. Ernst, P. T. Joo, K. U. Lao and R. A. DiStasio Jr., 2021, arXiv:2102.02354v1.

(207) B. Chan, A. T. B. Gilbert, P. M. W. Gill and L. Radom, *J. Chem. Theory Comput.*, 2014, **10**, 3777–3783.

(208) A. Karton, R. J. O'Reilly, B. Chan and L. Radom, *J. Chem. Theory Comput.*, 2012, **8**, 3128–3136.

(209) E. Romero-Montalvo and G. A. DiLabio, *J. Phys. Chem. A*, 2021, **125**, 3369–3377.

(210) V. Oliveira, E. Kraka and D. Cremer, *Phys. Chem. Chem. Phys.*, 2016, **18**, 33031–33046.

(211) N. Mehta, T. Fellowes, J. M. White and L. Goerigk, *J. Chem. Theory Comput.*, 2021, **17**, 2783–2806.

(212) V. M. Miriyala and J. Řezáč, *J. Phys. Chem. A*, 2018, **122**, 2801–2808.

(213) K. Kříž, M. Nováček and J. Řezáč, *J. Chem. Theory Comput.*, 2021, **17**, 1548–1561.

(214) K. U. Lao and J. M. Herbert, *J. Chem. Phys.*, 2013, **139**, 034107.

(215) K. U. Lao and J. M. Herbert, *J. Phys. Chem. A*, 2015, **119**, 235–252.

(216) N. Mardirossian, D. S. Lambrecht, L. McCaslin, S. S. Xantheas and M. Head-Gordon, *J. Chem. Theory Comput.*, 2013, **9**, 1368–1380.

(217) P. D. Mezei, G. I. Csonka, A. Ruzsinszky and J. Sun, *J. Chem. Theory Comput.*, 2015, **11**, 360–371.

(218) S. Zahn, D. R. Macfarlane and E. I. Izgorodina, *Phys. Chem. Chem. Phys.*, 2013, **15**, 13664–13675.

(219) R. Sure and S. Grimme, *J. Chem. Theory Comput.*, 2015, **11**, 3785–3801.

(220) D. I. Sharapa, J. T. Margraf, A. Hesselmann and T. Clark, *J. Chem. Theory Comput.*, 2017, **13**, 274–285.

(221) R. Sedlak, T. Janowski, M. Pitoňák, J. Řezáč, P. Pulay and P. Hobza, *J. Chem. Theory Comput.*, 2013, **9**, 3364–3374.

(222) J. Calbo, E. Ortí, J. C. Sancho-García and J. Aragó, *J. Chem. Theory Comput.*, 2015, **11**, 932–939.

(223) Z. Ni, Y. Guo, F. Neese, W. Li and S. Li, *J. Chem. Theory Comput.*, 2021, **17**, 756–766.

(224) H. Zhang, J. Krupa, M. Wierzejewska and M. Biczysko, *Phys. Chem. Chem. Phys.*, 2019, **21**, 8352–8364.

(225) K. N. Kirschner, W. Heiden and D. Reith, *ACS Omega*, 2018, **3**, 419–432.

(226) C. Greenwell and G. J. O. Beran, *Cryst. Growth Des.*, 2020, **20**, 4875–4881.

(227) S. L. J. Lahey, T. N. Thien Phuc and C. N. Rowley, *J. Chem. Inf. Model.*, 2020, **60**, 6258–6268.

(228) J. Řezáč, D. Bím, O. Gutten and L. Rulíšek, *J. Chem. Theory Comput.*, 2018, **14**, 1254–1266.

(229) The dcp package, https://github.com/aoterodelaroza/dcp, (accessed October 2021).

(230) The acpfit package, https://github.com/aoterodelaroza/acpfit, (accessed October 2021).

(231) AOR GitHub repository, https://github.com/aoterodelaroza, (accessed October 2021).

(232) M. R. Osborne, B. Presnell and B. A. Turlach, *IMA J. Numer. Anal.*, 2000, **20**, 389–403.

(233) M. Schmidt, PhD Thesis, University of British Columbia, 2010.

(234) Optimization Methods for L1-Regularization, https://www.cs.ubc.ca/tr/2009/tr-2009-19, (accessed March 8, 2021).

(235) M. J. Frisch *et al.*, Gaussian 16 (Revision B.01), Gaussian Inc., Wallingford, CT, 2016.

(236) F. Neese, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**, 73–78.

(237) N. Mardirossian and M. Head-Gordon, *Mol. Phys.*, 2017, **115**, 2315–2372.

(238) A. S. Mahadevi and G. N. Sastry, *Chem. Rev.*, 2016, **116**, 2775–2825.

(239) J. Hostaš and J. Řezáč, *J. Chem. Theory Comput.*, 2017, **13**, 3575–3585.

(240) V. K. Prasad, A. Otero-de-la-Roza and G. A. DiLabio, *Electron. Struc.*, 2021, **3**, 034007.

(241) J. C. Faver, M. L. Benson, X. He, B. P. Roberts, B. Wang, M. S. Marshall, C. D. Sherrill and K. M. Merz, *PLoS One*, 2011, **6**, 18868.

(242) M. Cutini, I. Bechis, M. Corno and P. Ugliengo, *J. Chem. Theory Comput.*, 2021, **17**, 2566–2574.

(243) Y. W. Hsiao, E. Sanchez-Garcia, M. Doerr and W. Thiel, *J. Phys. Chem. B*, 2010, **114**, 15413–15423.

(244) J. Antony and S. Grimme, *J. Comput. Chem.*, 2012, **33**, 1730–1739.

(245) M. Lepšík, J. Řezáč, M. Kolář, A. Pecina, P. Hobza and J. Fanfrlík, *ChemPlusChem*, 2013, **78**, 921–931.

(246) C. N. Cavasotto, in *Methods in Molecular Biology*, Humana Press Inc., 2020, vol. 2114, pp. 257–268.

(247) D. P. Harding, L. J. Kingsley, G. Spraggon and S. E. Wheeler, *J. Comput. Chem.*, 2020, **41**, 1175–1184.

(248) C. Zhang, S. Qin, B. Hu, J. Lv, Z. Yang, W. Yan, J. Wang, N. Huang and Z. Huang, *Nucleosides, Nucleotides and Nucleic Acids*, 2019, **38**, 567–577.

(249) A. C. Fogarty, E. Duboué-Dijon, F. Sterpone, J. T. Hynes and D. Laage, *Chem. Soc. Rev.*, 2013, **42**, 5672–5683.

(250) D. A. Britz and A. N. Khlobystov, *Chem. Soc. Rev.*, 2006, **35**, 637–659.

(251) D. R. Kauffman and A. Star, *Angew. Chemie Int. Ed.*, 2008, **47**, 6550–6570.

(252) D. Cao, X. Zhang, J. Chen, W. Wang and J. Yun, *J. Phys. Chem. B*, 2003, **107**, 13286–13292.

(253) M. S. Lohse and T. Bein, *Adv. Funct. Mater.*, 2018, **28**, 1705553.

(254) A. Germain and P. Ugliengo, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, 2020, vol. 12253 LNCS, pp. 745–753.

(255) B. Martínez-Bachs, S. Ferrero and A. Rimola, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland

GmbH, 2020, vol. 12251 LNCS, pp. 683–692.

(256) A. Rimola, S. Ferrero, A. Germain, M. Corno and P. Ugliengo, *Minerals*, 2021, **11**, 1–25.

(257) S. Ferrero, L. Zamirri, C. Ceccarelli, A. Witzel, A. Rimola and P. Ugliengo, *Astrophys. J.*, 2020, **904**, 11.

(258) E. F. Van Dishoeck, E. Herbst and D. A. Neufeld, *Chem. Rev.*, 2013, **113**, 9043–9085.

(259) A. L. Steber, C. Pérez, B. Temelso, G. C. Shields, A. M. Rijs, B. H. Pate, Z. Kisiel and M. Schnell, *J. Phys. Chem. Lett.*, 2017, **8**, 5744–5750.

(260) L. Ruiz Pestana, N. Mardirossian, M. Head-Gordon and T. Head-Gordon, *Chem. Sci.*, 2017, **8**, 3554–3565.

(261) A. P. Gaiduk, J. Gustafson, F. Gygi and G. Galli, *J. Phys. Chem. Lett.*, 2018, **9**, 3068–3073.

(262) R. A. DiStasio Jr., B. Santra, Z. Li, X. Wu and R. Car, *J. Chem. Phys.*, 2014, **141**, 084502.

(263) M. Chen, H.-Y. Ko, R. C. Remsing, M. F. Calegari Andrade, B. Santra, Z. Sun, A. Selloni, R. Car, M. L. Klein, J. P. Perdew and X. Wu, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 10846–10851.

(264) B. Hartke, Eur. *Phys. J. D*, 2003, **24**,57–60.

(265) F. F. Guimaräes, J. C. Belchior, R. L. Johnston and C. Roberts, *J. Chem. Phys.*, 2002, **116**, 8327–8333.

(266) A. Lenz and L. Ojamäe, *Phys. Chem. Chem. Phys.*, 2005, **7**, 1905–1911.

(267) H. J. Kulik, N. Luehr, I. S. Ufimtsev and T. J. Martinez, *J. Phys. Chem. B*, 2012, **116**, 12501–12509.

(268) S. Schmitz, J. Seibert, K. Ostermeir, A. Hansen, A. H. Göller and S. Grimme, *J. Phys. Chem. B*, 2020, **124**, 3636–3646.

(269) B. Vorlová, D. Nachtigallová, J. Jirásková-Vaníčková, H. Ajani, P. Jansa, J. Řezáč, J. Fanfrlík, M. Otyepka, P. Hobza, J. Konvalinka and M. Lepšík, *Eur. J. Med. Chem.*, 2015, **89**, 189–197.

(270) V. M. Miriyala and J. Řezáč, *J. Phys. Chem. A*, 2018, **122**, 2801–2808.

(271) J. Zheng, X. Xu and D. G. Truhlar, *Theor. Chem. Acc.*, 2011, **128**, 295–305.

(272) C. Riplinger, B. Sandhoefer, A. Hansen and F. Neese, *J. Chem. Phys.*, 2013, **139**, 134101.

(273) P. Morgante and R. Peverati, *Chem. Phys. Lett.*, 2021, **765**, 138281.

(274) W. Kabsch, *Acta Crystallogr. Sect. A*, 1976, **32**, 922–923.

(275) B. Boekfa, S. Choomwattana, P. Khongpracha and J. Limtrakul, *Langmuir*, 2009, **25**, 12990–12999.

(276) S. Kim, D. J. Robichaud, G. T. Beckham, R. S. Paton and M. R. Nimlos, *J. Phys. Chem. A*, 2015, **119**, 3604–3614.

Chapter 9

Small basis set density-functional theory methods corrected with atom-centered potentials

Abstract

Density-functional theory (DFT) is currently the most popular method for modeling non-covalent interactions and thermochemistry. The accurate calculation of non-covalent interaction energies, reaction energies, and barrier heights requires choosing an appropriate functional and, typically, a relatively large basis set. Deficiencies of the density-functional approximation and the use of limited basis set are the leading sources of error in the calculation of non-covalent and thermochemical properties in molecular systems. In this article, we present three new DFT methods based on the BLYP, M062X and CAM-B3LYP functionals in combination with the 6-31G* basis set and corrected with atom-centered potentials (ACPs). ACPs are one-electron potentials that have the same form as effective-core potentials, except they do not replace any electrons. The ACPs developed in this work are used to generate energy corrections to the underlying DFT/basis-set method such that the errors in predicted chemical properties are minimized while maintaining the low computational cost of the parent methods. ACPs were developed for the elements H, B, C, N, O, F, Si, P, S, Cl. The ACP parameters were determined using an extensive training set of 118,655 data points, mostly of complete basis set coupled-cluster level quality. The target molecular properties for the ACP-corrected methods include non-covalent interaction energies, molecular conformational energies, reaction energies, barrier heights, and bond separation energies. The ACPs were tested first on the training set and then on a validation set of 42,567 additional data points. We show that the ACP-corrected methods can predict the target molecular properties with accuracy close to complete basis set wavefunction theory methods, but at a computational cost of a double-ζ DFT methods. This makes the new BLYP/6-31G*-ACP, M062X/6-31G*-ACP, and CAM-B3LYP/6-31G*-ACP methods uniquely suited to the calculation of non-covalent, thermochemical, and kinetic properties in large molecular systems.

1. Introduction

Over the past few decades, density functional theory (DFT) has become the leading approach in the quantum mechanical (QM) modeling of various molecular properties.[1] DFT's success can be attributed to its favorable balance between computational cost and accuracy, along with the existence of efficient algorithmic implementations widely available in modern software packages. One of the main sources of error in DFT is the choice of exchange-correlation functional approximation, which determines the accuracy of a given DFT method for a particular purpose. For instance, it is known that common density functional approximations are unable to accurately describe dispersion forces, which are critical when modeling non-covalent interactions and chemical reactions involving large molecules.[2–4] As a result, several works have focused on improving the accuracy of common density functional approximations by developing various dispersion correction techniques.[5–8]

It has been shown that dispersion-corrected DFT methods in combination with large basis sets can predict various molecular properties and in particular non-covalent interaction energies of medium-size systems with accuracy similar or slightly lower than that obtained with nearly complete basis set wavefunction theory methods.[9–14] However, even with recent advances in computer technology, the applicability of dispersion-corrected DFT methods with a large basis set is a challenge for systems containing more than ca. one hundred atoms. This is unfortunate because there are many interesting problems involving systems in this molecular size range: supra-molecular and (bio)chemical complexes, nanostructured materials and surfaces, enzyme active sites, and many more.[15–17] The reason for this limitation is the unfavorable increase in computational cost as approximately the third power in the number of basis functions for common DFT methods. Consequently, the development of computationally inexpensive DFT based methods that allow efficient and accurate modeling of large systems is an important area of research.[18–21]

The reduction in computational cost of dispersion-corrected DFT with the use of small double-ζ basis set, such as 6-31G*, allows for the modeling of large molecular systems. The primary sources of error in dispersion-corrected DFT plus 6-31G* are the exchange-correlation functional approximation and basis set incompleteness error caused by the small size of the basis set. Methods have been recently proposed that show that these shortcomings can be mitigated efficiently.[22–24] For instance, Grimme and co-workers have shown that the D3[25,26] dispersion correction when combined with two additional semi-empirical corrections[27,28] designed to mitigate basis set incompleteness error (collectively known as the "3c" approach[28]), alleviate the deficiencies of the PBEh, HSE, and B3LYP functionals with a double-ζ basis set (yielding PBEh-3c[29], HSE-3c[30], and B3LYP-3c[31] methods). The 3c approach, which was initially proposed for a minimal basis set Hartree–Fock (HF) method (HF-3c[28]), has also recently been extended to triple-ζ basis set DFT methods (B97-3c[32] and r^{2}SCAN-3c[33]). Several other methodologies based on DFT have been proposed in the literature[34–40], and this underscores the interest in developing computationally inexpensive DFT based methods for modeling large molecular systems.

We have shown in earlier works that atom-centered potentials[41] (ACPs) offer a useful way to mitigate the underlying shortcomings of HF and DFT methods.[42–55] ACPs are similar to one-electron effective-core potentials[56,57] (ECPs) in functional form, except ACPs do not replace any electrons. Sharing a form similar to ECPs allows ACPs to be used in any software package that implements the use of ECPs. ACPs can be developed to yield energy corrections that minimize the errors in predicted properties for a target method and basis set combination by parametrization against high-level reference data. In this way, ACPs can be used to efficiently mitigate the shortcomings of double-ζ basis set DFT methods, since the use of ACPs incurs only ca. 10% increase[58] in the computational cost relative to the uncorrected method.

In our recent work[58], we developed four sets of ACPs for ten elements in combination with small and minimal basis set HF. These ACPs were trained against a set of 73,832 non-covalent properties (interaction energies, molecular conformational energies, and molecular deformation energies). Only non-covalent properties were used in the training set because small basis set HF, and HF in general, is limited to applications that do not involve bond breaking or formation, such as fast geometry optimizations, high-throughput conformer screening, and prediction of non-covalent interaction strengths in large systems. Small basis set DFT is a much more promising approach for modeling thermochemical, kinetic, as well as non-covalent properties since, as previously mentioned, the parent DFT methods are already reasonably successful in the calculation of these quantities. Therefore, ACPs developed for double-ζ basis set DFT methods that mitigate the errors associated with the density functional approximation error and basis set incompleteness error can be used to model all the aforementioned non-covalent properties as well as reaction energies, transition state searches, and in barrier height calculations of large molecular systems.

In this work, we developed ACPs for three density functionals combined with a double-ζ basis set (6-31G*[59,60]) and Grimme's D3[25,26] dispersion correction scheme where applicable. The functionals chosen were BLYP[61,62] (generalized-gradient approximation or GGA functional), M062X[63] (hybrid meta-GGA functional), and CAM-B3LYP[64] (range-separated hybrid functional), mainly due to their popularity and performance for the target properties.[13,14,22,65–67] The main target molecular properties are non-covalent interaction energies, molecular conformational energies, reaction energies, barrier heights, and bond separation energies. ACPs were developed for ten elements commonly encountered in organic chemistry and biochemistry (H, C, N, O, F, P, S, Cl) plus boron and silicon. The ACP development was carried out using a training set composed of 118,655 data points calculated at a high level of theory. We used a regularized linear least-squares fitting procedure (the LASSO[68–70] regression method) to obtain the parameters of the ACPs, which greatly simplifies the use of such a large training set. The strengths and weaknesses of the developed ACPs are evaluated and discussed based on their performance on the training set and a validation set consisting of additional 42,567 data points.

2. Computational Details

The ACP development procedure employed in this article has been described in detail in our earlier works[42–44,58]. We summarize it here for convenience. The mathematical form of an ACP is:

$$\hat{V}_{ACP} = \sum_{\alpha}\left(V^{\alpha}_{local}(r) + \sum_{l=0}^{L-1}\sum_{m=-l}^{l} \delta V_l^{\alpha}(r)\, |Y_{lm}\rangle\langle Y_{lm}|\right) \tag{1}$$

where $\delta V_l^{\alpha}(r) = V_l^{\alpha}(r) - V_{local}^{\alpha}(r)$, α represents atom, r is the distance, and $|Y_{lm}\rangle\langle Y_{lm}|$ are projection operators using real spherical harmonics based on atom α with l angular momentum quantum numbers and m magnetic quantum numbers. The $V_{local}^{\alpha}(r)$ and $\delta V_l^{\alpha}(r)$ terms in Equation 1 are represented by N Gaussian-type functions:

$$V_l^{\alpha}(r) = \sum_{n=1}^{N} c_{ln}^{\alpha} \exp(-\xi_{ln}^{\alpha} r^2) \qquad \text{for} \quad l = 0, 1, 2, \ldots, L \tag{2}$$

where the coefficients (c_{ln}^{α}) and exponents (ξ_{ln}^{α}) are adjustable parameters determined via a regularized least-squares fit to reference data during the ACP development.

In order to find the exponents and coefficients that best mitigate the errors in the properties predicted using the target method and basis set combination, the ACP operator (Equation 1) is first added as a perturbative correction to the Hamiltonian. To first order in the ACP perturbation, the energy correction is:

$$E_{ACP}(\{c_{ln}^{\alpha}\}, \{\xi_{ln}^{\alpha}\}) = \sum_{i} \langle \psi_i | \hat{V}_{ACP} | \psi_i \rangle \tag{3}$$

$$E_{ACP}(\{c_{ln}^{\alpha}\}, \{\xi_{ln}^{\alpha}\}) = \sum_{\alpha l n} c_{ln}^{\alpha} \sum_{i} \langle \psi_i | (|Y_{lm}\rangle \exp(-\xi_{ln}^{\alpha} r^2) \langle Y_{lm}|) | \psi_i \rangle \tag{4}$$

$$E_{ACP}(\boldsymbol{c}, \boldsymbol{\xi}) = \sum_{\alpha l n} c_{ln}^{\alpha}\, \Delta E_{ln}^{\alpha}(\xi_{ln}^{\alpha}) = \boldsymbol{c} \cdot \boldsymbol{\Delta}\mathbf{E}(\boldsymbol{\xi})^{T} \tag{5}$$

where the index i in Equations 3 and 4 runs over occupied molecular orbitals ψ.

The $\Delta E_{ln}^{\alpha}(\xi_{ln}^{\alpha}) = \langle \psi_i | (|Y_{lm}\rangle \exp(-\xi_{ln}^{\alpha} r^2) \langle Y_{lm}|) | \psi_i \rangle$ integrals (Equations 4 and 5) are the ACP energy terms and they are equal to the difference between the energy when an ACP with exponent ξ_{ln}^{α} is applied and the energy in absence of any ACP, divided by the ACP coefficient. In matrix notation, the $\boldsymbol{c}$

and $\mathbf{\Delta E}(\boldsymbol{\xi})^T$are vectors of ACP coefficients and ACP energy terms, respectively. Equations 1–5 indicate that the energy corrections introduced by the ACPs decay exponentially with interatomic distances and depend on the molecular wavefunction as well as the local chemical environment of a given atom.

The ACP development process starts with the compilation of a training set of target molecular properties. The training set must be chemically diverse and composed of systems exclusively with atoms for which the ACPs are being developed. Next, a set of exponents (ξ_{ln}^{α}) and angular momenta (l) on each atom are selected, and the ACP energy terms ($\Delta E_{ln}^{\alpha}(\xi_{ln}^{\alpha})$) are computed for each exponent and each entry in the training set. Once all the ACP energy terms are calculated, the optimal ACP coefficients c_{ln}^{α} and the associated exponents are determined using a regularized least-squares fitting subject to a constraint on the sum of the absolute values of the coefficients. The ACP development process ends with the generation of the ACPs for the target method and basis set combination.

The ACPs in this work are designed for correcting the BLYP, M062X, and CAM-B3LYP functionals in combination with the 6-31G* basis set. These ACPs are tied to these method and basis set combinations and are not transferable to other methods. The target elements chosen for ACP development are H, B, C, N, O, F, Si, P, S, and Cl; most of these atoms are common in organic chemistry and biochemistry applications. Twenty-nine exponents (ξ_{ln}^{α}) were considered: 0.12 to 0.30 in 0.02 steps, 0.40 to 2.00 in 0.10 steps, and 2.50 to 3.00 in 0.50 steps. Angular momenta for the ACP energy terms ($\Delta E_{ln}^{\alpha}(\xi_{ln}^{\alpha})$) were used up to the maximum angular momentum of the valence orbital basis functions present in the 6-31G* basis set: up to *s* for H and *d* for B, C, N, O, F, Si, P, S, and Cl. The total number of ACP energy terms was 1,102. This way of generating the ACP energy terms and carrying out the fitting procedure is identical to our previous works[42–44,55,58]. LASSO (Least Absolute Shrinkage and Selection Operator) regression[68–70] is employed to solve the regularized least-squares fitting problem. The advantage of LASSO is that it automatically selects the best subset of ACP energy terms and discards the others by assigning a zero coefficient to them. All the single-point energy calculations were performed with the *Gaussian16*[71] software package. The D3 parameters used are listed in the SI, along with the ACP coefficients and exponents for each method. An example of the usage of ACPs in the *Gaussian16* program is also given in the SI.

The training set (Table 1) used to parameterize the ACPs comprises data sets from the literature that represent non-covalent and covalent properties such as interaction energies, molecular conformational energies, reaction energies, barrier heights, and bond separation energies. This choice of training set properties is motivated by the potential target applications of the ACP-corrected small basis set DFT methods, namely fast geometry optimizations, conformer screening, and modeling of chemical reactions and non-covalent interaction strengths of large systems. The training set comprises 19,439 non-covalent interaction energies, 11,161 molecular conformational energies, 8,315 reaction energies, 58,197 barrier heights, and 4,502 bond separation energies. Our training set also includes 240 molecular isomerization energies, 219 total atomization energies, and 16,582 molecular deformation energies. In addition to the training set, we also assembled a validation set (Table 2), a collection of data used to test the accuracy of properties computed using ACPs for systems not included in the training set. In total, the validation set consists of 27,783 non-covalent interaction energies, 9,491 molecular conformational energies, 5,205 reaction energies, and 88 barrier heights. The structures and reference energies of all data points in the training and validation sets are provided in the SI. Most of the reference data used in both the training and validation sets were calculated with nearly complete basis set wavefunction theory methods, with any exception noted in Tables 1 and 2.

Table 1. List of data sets used for training the ACPs.

Data set(s)	Data points	Range of reference data (in kcal/mol)	Description of data points
Non-covalent interaction energies of molecular complexes*[a]*:			
HBC6[77,78], MiriyalaHB104[79,80], IonicHB[81], HB375x10[82], IHB100x10[82], HB300SPXx10[83], CARBH12[14]	6,409	-37.01 to +16.30	hydrogen bonding interactions
S22x5[78,84,85], S66x8[86–88], S66a8[87], A21x12[89–91], NBC10ext[78,92–95], 3B-69-DIM[96], 3B-69-TRIM[96], HW30[97]	1,895	-35.76 to +9.34	mix character non-covalent interactions
B-set[b,64], F-set[b,64], Si-set[b,64], P-set[b,64], S-set[b,64], Cl-set[b,64], Sulfurx8[98]	1,000	-68.05 to +21.57	monomers containing at least one B, F, Si, P, S, and Cl atom
Pisub[b,99,100], Pi29n[101], BzDC215[102], C2H4NT[95]	379	-18.30 to +10.33	non-stacked and stacked π-π interactions
Hill18[103], X40x10[104]	238	-14.14 to +11.95	halogen bonding interactions
PNICO23[14,105]	23	-10.97 to -0.64	pnicogen bonding interactions
ADIM6[14,25,106], HC12[107]	18	-5.60 to -1.30	hydrophobic interactions
BBI[108], SSI[108], NucTAA[b,c,109–112], CarbhydBz[113], CarbhydNaph[114], CarbhydAroAA[b,115], CarbhydAro[b,116], WatAA[b,117], HSG[78,118], PLF547[119], JSCH[84], DNAstack[120], DNA2body[120], ACHC[121], BDNA[122], NucBTrimer[b,123]	4,756	-100.86 to +64.19	non-covalent interactions present in various biomolecules
Water38[124], Water1888[95,125–127], Water-2body[d,54]	2,336	-92.89 to +5.10	hydrogen-bonded water dimers and $(H_2O)_n$ clusters where n=3–10
CH4PAH[128,129], CO2MOF[130], CO2PAH[131], CO2NPHAC[132], BzGas[133]	876	-6.02 to +12.17	non-covalent interactions between gas and substrate molecules
SSI-anionic[c,108], WatAA-anionic[b,c,117], HSG-anionic[c,78,118], PLF547-anionic[c,119], IonicHB-anionic[c,81], IHB100x10-anionic[c,82], Ionic43-anionic[c,134]	1,509	-135.11 to +88.94	anionic interactions
Molecular conformational energies*[e]*:			
37Conf8[135], DCONF[136], ICONF[14], MCONF[137], Torsion21[138], MolCONF[139]	8,280	+0.0005 to +25.06	various molecules representing pharmaceuticals, catalysts, synthetic

Data set(s)	Data points	Range of reference data (in kcal/mol)	Description of data points
			precursors, industrial chemicals, and organic compounds
PEPCONF-Dipeptide[b,140], TPCONF[141], P76[142], YMPJ[143], SPS[f,144], rSPS[f,145], UpU46[f,146], SCONF[14,147], DSCONF[148], SacchCONF[149], CCONF[150]	2,082	-4.09 to +19.74	molecules representative of proteins, DNA, RNA, and carbohydrates
ACONF[151], BCONF[152], PentCONF[153]	421	+0.14 to +16.66	hydrocarbon-like molecules
Undecamer125[154]	124	+0.06 to +1.87	$(H_2O)_{11}$ clusters
PEPCONF-Dipeptide-anionic[b,f,140], MolCONF-anionic[f,139]	254	-0.47 to +10.96	negatively charged molecules
Reaction energies[g]:			
MN-RE[155]	7,555	-217.97 to +242.47	automatically generated reactions using molecules from Minnesota Database2015B[156]
BH9-RE[74]	449	-89.85 to +116.88	from BH9 set comprising chemical reactions belonging to nine types common in organic chemistry and biochemistry
DIE60[157]	60	-6.14 to +8.60	double-bond migration reactions in conjugated dienes
FH51[158,159]	51	-150.81 to -0.18	reactions involving various organic and inorganic molecules
BSR36[160,161]	36	+2.24 to +49.82	hydrocarbon bond separation reactions
BH76RC[162–164]	30	-103.91 to +5.60	hydrogen and non-hydrogen atom transfer reactions of small molecules
G2RC[14,164,165]	23	-154.04 to -2.18	reactions whose reactants and products had been taken from the G2/97 set
RC21[14]	21	-6.72 to +126.56	organic radical fragmentation and rearrangement reactions
CR20[166]	20	-35.70 to -7.66	cyclo-reversion reactions
PlatonicHD6[167], PlatonicID6[167], PlatonicIG6[167]	18	-43.64 to +501.85	homodesmotic, isodesmic, and isogyric reactions involving platonic hydrocarbon cages, C_nH_n (where n = 4,6,8,10,12,20)
AlkIsod14[168]	14	+2.20 to +15.40	isodesmic reactions involving C_nH_{2n+2} alkanes (where n=3-8)
DARC[14,164,169]	14	-60.80 to -14.00	Diels-Alder reactions
DC13[14,63,178,179,170–177]	12	-106.00 to +152.60	reactions that were known to be difficult for DFT methods
WCPT6[180]	6	-0.86 to +11.14	tautomeric water-catalyzed proton transfer reactions
NBPRC[161,164,181]	6	-31.20 to +40.40	reactions involving NH_3/BH_3 and PH_3/BH_3
Barrier height energies[h]:			
Grambow2020-B97D3[i,j,182]	32,722	-44.42 to +221.34	reactions involving H, C, N, and O generated using automated potential energy surface exploration
Grambow2020-ωB97XD3[i,j,182]	23,922	-15.70 to +201.33	reactions involving H, C, N, and O generated using automated potential energy surface exploration
BH9[k,74]	898	-96.26 to +144.39	chemical reactions belonging to nine types common in organic chemistry and biochemistry
E2SN2[i,k,183]	418	-15.54 to +50.78	competing E2 and S_N2 reactions

Data set(s)	Data points	Range of reference data (in kcal/mol)	Description of data points
HTBH38[k,163]	38	+1.70 t0 +38.40	hydrogen atom transfer reactions of small molecules
NHTBH38[k,162]	38	-12.54 to +106.18	non-hydrogen atom transfer reactions of small molecules
WCPT27[k,180]	27	-6.38 to +81.24	water-catalyzed proton-transfer reactions
BHROT27[k,14]	27	+1.01 to +17.24	rotation around single bonds
BHPERI26[k,164,184]	26	+0.50 to +39.70	pericyclic reactions
DBH24[k,185,186]	24	-2.40 to +82.14	diverse reactions involving small molecules
INV24[k,187]	24	+4.10 to +79.70	inversion and racemization reactions
CRBH20[k,188]	20	+33.71 to +52.42	cyclo-reversion reactions of heterocyclic rings
PX13[k,189,190]	13	-29.97 to +56.19	proton exchange reactions in small clusters of H_2O, NH_3, and HF
Bond separation energies[l]:			
BSE49[76]	4,502	9.38 to 177.24	breaking of 49 unique X-Y type single bonds (except H-H, H-F, and H-Cl) into corresponding radical fragments, where X and Y are H, B, C, N, O, F, Si, P, S, Cl
Others[m]:			
MOLdef[b,43], MOLdef-H2O[d,191,192], MOLdef-TS[n,o,74]	16,582	-98.43 to +49.38	molecular deformation energies of various molecules deformed along their normal modes
ISO34[14,193], ISOL24[14,194], IDISP[14,161,164,193,195,196], EIE22[197], PArel[14], AlkIsomer11[168], PAH6[198], Styrene45[170], TAUT15[14], H2O16Rel5[199], H2O20Rel10[200], SW49Rel6[201], SW49Rel345[201]	240	-60.28 to +124.46	isomerization energies
W4-17[202], PlatonicTAE6[167], AlkAtom19[168]	219	+2484.26 to +4621.46	total atomization energies

[a] defined as the difference between the energy of the complex and the sum of the monomer energies. A negative interaction energy indicates the complex is more stable than the separated monomers.

[b] the reference data was recalculated using DLPNO-CCSD(T)/CBS (see Reference 58).

[c] comprises non-covalently bound dimer complexes where at least one of the monomers is negatively charged.

[d] the reference data was calculated at the CCSD(T)/CBS level using the same extrapolation method as in Reference 124.

[e] defined as the difference between the energy of a particular conformer and a lower-energy conformer of the same molecule.

[f] comprises negatively charged conformers.

[g] defined as the difference between the sum of energies of reactants minus that of products.

[h] forward barrier height is the energy difference between transition state and reactant(s) or pre-reaction complex; reverse barrier height is the energy difference between transition state and product(s) or post-reaction complex.

[i] the reference data was recalculated using DLPNO-CCSD(T)/CBS with the same extrapolation method as in Reference 74.

[j] the barrier heights are relative to the pre- or post-reaction complexes for forward and reverse barriers, respectively.

[k] the barrier heights are relative to the isolated reactant(s) or product(s) for forward and reverse barriers, respectively.

[l] defined as the difference between the energy of a molecule and its radical fragments formed by cleavage of a particular bond.

[m] includes molecular deformation energies, isomerization energies, and total atomization energies. Molecular deformation energy is the difference between the energy of a molecule deformed along a particular normal mode and the energy of the same molecule at equilibrium. Isomerization energy is the energy difference between a molecule and one of its isomers. Total atomization energy is the energy difference between a molecule and the sum of the energies of all its constituent atoms.

[n] includes deformations along the imaginary mode of transition state structures from Reference 74. The reference data was calculated at DLPNO-CCSD(T)/CBS level using the same extrapolation method as used in Reference 58.

[o] contains reference data which is negative in magnitude due to deformation along the imaginary normal mode, indicating that the deformed molecule is more stable than the transition state structure.

Table 2. List of data sets used for validating the ACPs.

Data set(s)	Data points	Range of reference data (in kcal/mol)	Description of data points
Non-covalent interaction energies of molecular complexes*[a]*:			
BlindNCI[203], DES15K[204], NENCI-2021[205]	17,413	-33.78 to +186.83	mix character non-covalent interactions
CE20[189,190], WaterOrg[206]	2,396	-46.58 to -10.76	hydrogen bonding interactions
R160x6[207], R739x5[208]	5,290	-12.02 to +6.79	close contact interactions
CHAL336[209]	48	-30.85 to -1.57	chalcogen bonding interactions
XB45[210]	33	-13.11 to -0.89	halogen bonding interactions
L7[211,212], S12L[9,11,212], S30L[213], Ni2021[214]	54	-416.08 to -1.68	large molecules relevant in supramolecular chemistry and biochemistry
C60dimer[215]	14	-6.88 to +12.07	C_{60} dimers
H2O20Bind10[200]	10	-200.54 to -196.59	$(H_2O)_{20}$ clusters
HW6Cl[b,200,216], HW6F[b,200,216], FmH2O10[b,200,216], SW49Bind345[b,201], SW49Bind6[b,201], Anionpi[b,217], IL236[b,218], DES15K-anionic[b,204], NENCI-2021-anionic[b,205], CHAL336-anionic[b,209], XB45-anionic[b,210], S30L-anionic[b,213]	2,525	-171.42 to +66.15	anionic interactions
Molecular conformational energies*[c]*:			
SafroleCONF[219], AlcoholCONF[220], BeranCONF[221], Torsion30[d,222], ANI1ccxCONF[e,f,223]	7,447	+1E-3 to +49.96	Safrole or 5-(2-propenyl)-1,3-benzodioxol), small alcohol molecules, biaryl drug-like molecules, and small organic molecules,
MPCONF196[g,224], PEPCONF-Tripeptide[h,140], PEPCONF-Disulfide[140], PEPCONF-Cyclic[140], PEPCONF-Bioactive[140]	1,874	-0.47 to +81.00	peptide-like molecules
PEPCONF-Disulfide-anionic[i,140], PEPCONF-Bioactive-anionic[i,140]	170	+0.17 to +33.79	negatively charged molecules
Reaction energies*[j]*:			
W4-17-RE[f,155]	5,205	-380.97 to +364.92	automatically generated reactions using molecules from the W4-17[202] set
Barrier height energies*[k]*:			
WaterOrgBH[l,m]	88	+12.81 to +61.50	pericyclic reactions in absence and presence of water clusters

[a] defined as the difference between the energy of the complex and the sum of energy of the monomers. A negative interaction energy indicates the complex is more stable than the separated monomers.

[b] comprises non-covalently bound complexes with at least one negatively charged monomer.

[c] defined as the difference between the energy of a particular conformer and a lower-energy conformer of the same molecule.

[d] only 30 systems used; we could not find the rest of the systems mentioned in the supporting information of Reference 222.

[e] contains mostly conformational energies but also some molecular deformation energies.

[f] only a subset of the actual data used.

[g] only macrocyclic peptides used.

[h] only a subset from the PEPCONF[140] database for which reference data was recalculated at the DLPNO-CCSD(T)/CBS level of theory (see Reference 58 for more details).

[i] comprises negatively charged conformers.

[j] defined as the difference between the sum of energies of reactants minus that of products.

[k] forward barrier height is the energy difference between transition state and reactant(s) or pre-reaction complex; reverse barrier height is the energy difference between transition state and product(s) or post-reaction complex.

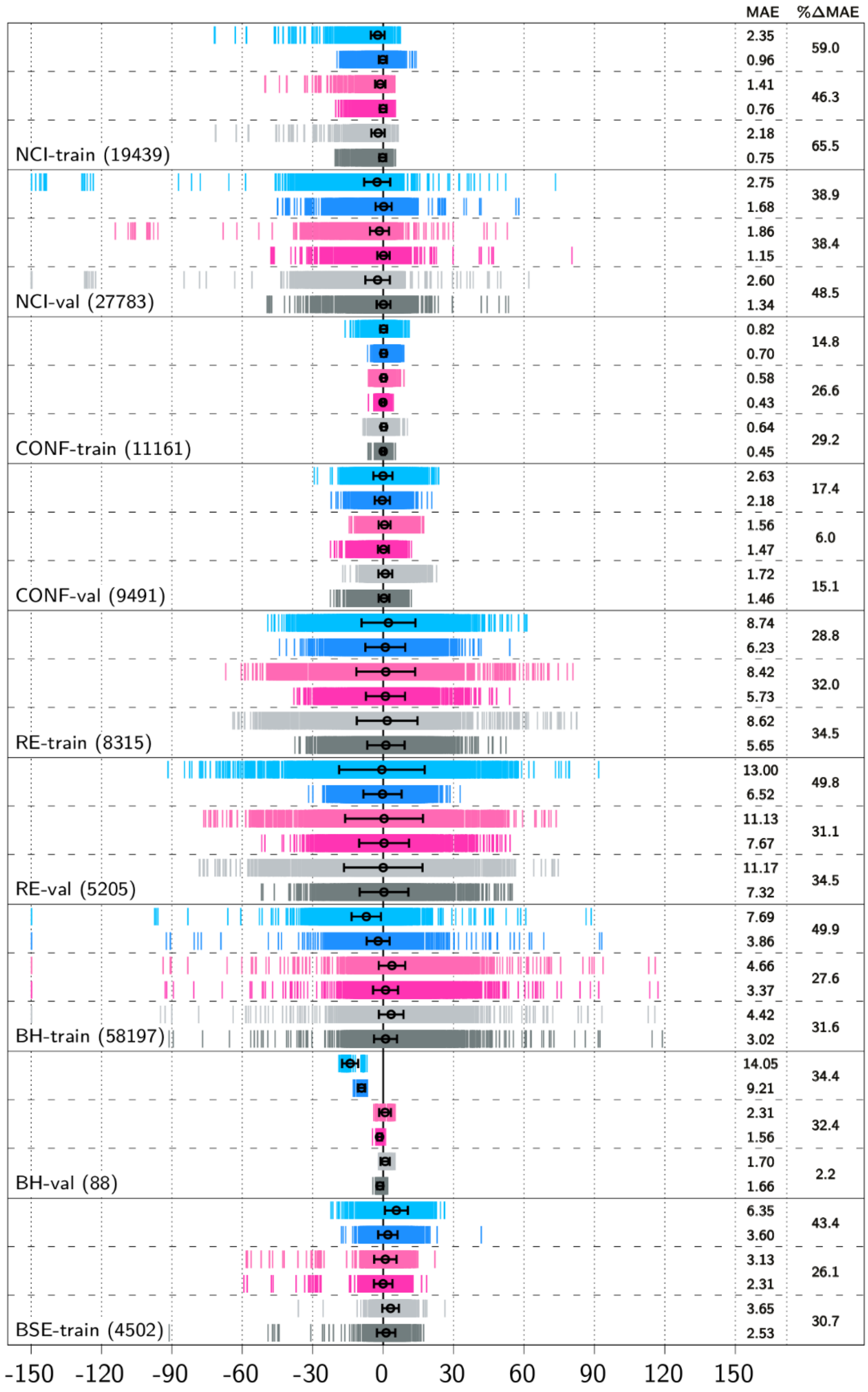
[l] the barrier heights are relative to the pre- or post-reaction complexes for forward and reverse barriers, respectively.

[m] contains unpublished data generated in an ongoing project, which will be published elsewhere.

3. Results and Discussion

Results obtained using the target methodology (BLYP-D3/6-31G*, M062X/6-31G*, and CAM-B3LYP-D3/6-31G*) with and without the proposed ACPs compared against the reference data are shown in Figure 1. The figure depicts the signed error distribution as vertical lines along with the mean signed errors (MSEs) (open circles) and the standard deviations (SDs) of the errors (horizontal black lines). The mean absolute errors (MAEs) of each method and the percentage changes in mean absolute error (%ΔMAE) on application of the ACPs are also given on the right. A more detailed comparison can be found in the supporting information (Tables S1 and S2 of the SI). A detailed breakdown of the error analysis of each method by subset can be found in Figures S1 and S2 and Tables S3 and S4 of the supporting information. In the following, the results obtained from the application of ACPs are discussed for the different molecular properties in the training and validation sets.

Figure 1. Error distribution (in kcal/mol) associated with the uncorrected and ACP-corrected double-ζ DFT methods. The various molecular properties represented are: "NCI" or non-covalent interaction energies, "CONF" or molecular conformational energies, "RE" or reaction energies, "BH" or barrier height energies, and "BSE" or bond separation energies. Suffixes "train" and "val" are short for training and validation, respectively (see Table 1 and 2). Methods shown include BLYP-D3/6-31G* (light blue), BLYP-D3/6-31G*-ACP (blue), M06-2X/6-31G* (light pink), M06-2X/6-31G*-ACP (pink), CAM-B3LYP-D3/6-31G* (light grey), and CAM-B3LYP-D3/6-31G*-ACP (grey). The black circles represent the mean signed errors (MSEs, kcal/mol) and the black error bars are the standard deviations of the error (SDs, kcal/mol). The numbers on the right-hand side of each panel are the mean absolute errors (MAEs, kcal/mol) and the percentage change in MAEs upon the application of ACPs (%ΔMAE) for each method. %ΔMAE is defined as [MAE(base method) – MAE(ACP-corrected method)] / MAE(base method) x 100%. The X-axis has been capped at -150 (left) and +150 kcal/mol (right) for clarity.


MAE
%ΔMAE
2.35
0.96
59.0
1.41
0.76
46.3
2.18
0.75
65.5
NCI-train (19439)
2.75
1.68
38.9
1.86
1.15
38.4
2.60
1.34
48.5
NCI-val (27783)
0.82
0.70
14.8
0.58
0.43
26.6
0.64
0.45
29.2
CONF-train (11161)
2.63
2.18
17.4
1.56
1.47
6.0
1.72
1.46
15.1
CONF-val (9491)
8.74
6.23
28.8
8.42
5.73
32.0
8.62
5.65
34.5
RE-train (8315)
13.00
6.52
49.8
11.13
7.67
31.1
11.17
7.32
34.5
RE-val (5205)
7.69
3.86
49.9
4.66
3.37
27.6
4.42
3.02
31.6
BH-train (58197)
14.05
9.21
34.4
2.31
1.56
32.4
1.70
1.66
2.2
BH-val (88)
6.35
3.60
43.4
3.13
2.31
26.1
3.65
2.53
30.7
BSE-train (4502)
-150
-120
-90
-60
-30
0
30
60
90
120
150

(i) Non-covalent interaction energies

Regarding non-covalent interaction energies ("NCI"), the application of ACPs to the double-ζ basis set DFT methods decreases the overall MAE by 46–65%, indicating a substantial improvement in the description of non-covalent interactions. This is reasonable given that ACPs are capable of mitigating the basis set incompleteness error which is known to greatly affect the calculation of non-covalent interaction energies.[43,44,72] The best methods based on the overall performance for non-covalent interactions are M062X/6-31G*-ACP and CAM-B3LYP-D3/6-31G*-ACP, both with an MAE of about 0.75 kcal/mol. For comparison, BLYP-D3/6-31G*-ACP yields an overall MAE of 0.96 kcal/mol. A closer look into individual NCI data sets (Figure S1 of SI) reveals that the ACPs applied to CAM-B3LYP-D3/6-31G* result in a more uniform reduction in the MAEs of the uncorrected method compared to M062X/6-31G*. BLYP-D3/6-31G*-ACP also performs uniformly better than the parent uncorrected method, although the individual data set MAEs are slightly higher than CAM-B3LYP-D3/6-31G*-ACP.

We now take a more detailed view at the performance of CAM-B3LYP-D3/6-31G*-ACP on the various non-covalent interaction types in the training set. Application of the ACPs to this functional result in MAE reductions for the hydrogen bonding and mixed-character non-covalent interactions greater than 50%. The description of halogen bonding and pnicogen bonding is also improved with MAE reductions in the range of about 19–57%, even though the number of data points for these interaction types in the training set is comparatively smaller than the others. Stacked and non-stacked π-π interactions are also described better with the ACPs, with MAEs reduced by about 17–34%. In addition to the various interaction types, ACPs were also trained on typical systems relevant in organic chemistry and biochemistry, where these non-covalent interactions operate co-operatively. Examples include interacting nucleotides or proteins interacting with carbohydrates, nucleotides, drugs, water, and other proteins. For CAM-B3LYP-D3/6-31G*, the MAE reduction caused by the ACPs in the data sets of biochemical significance range between 29% and 94%, indicating that ACPs significantly enhance the performance of CAM-B3LYP-D3/6-31G* for modeling non-covalent interactions in biomolecular systems. ACPs for CAM-B3LYP-D3/6-31G* also lead to a better description of other non-covalently interacting systems in the training set, including gas-substrate and water-water complexes, with MAE reductions in the 24–91% range.

Even though CAM-B3LYP-D3/6-31G*-ACP improves on the base uncorrected method for almost all non-covalent interaction types, some outliers with relatively high error exist, which is expected given the enormous size of the training set. There are only two NCI subsets in the training set (ADIM6 and

HC12) where CAM-B3LYP-D3/6-31G*-ACP yields higher MAEs than uncorrected CAM-B3LYP-D3/6-31G*, both featuring mainly hydrocarbon interactions and contributing only 18 data points to the training set. The increase in MAE upon application of ACPs is likely the result of the relative scarcity of pure hydrophobic type interactions in the training set. Besides hydrophobic contacts, anionic interactions are set of interaction types (SSI-anionic, WatAA-anionic, HSG-anionic, PLF547-anionic, IonicHB-anionic, IHB100x10-anionic, and Ionic43-anionic) where there is room for improvement. In this case, ACPs lead to an improved description compared to the uncorrected double-ζ basis set DFT methods. However, the remaining relatively high errors are probably due to the fact that the 6-31G* basis set lacks diffuse basis functions required, which are known to be required for modeling anionic systems.[73]

Given the overall good performance of ACPs for non-covalent interactions in the training set, we now examine their performance for systems outside the training set. Figure 1 shows that ACPs successfully bring down the overall MAE of the double-ζ basis set DFT methods by about 38–48% for the NCI validation subset (“NCI-val”). A more detailed look into the results (Figure S3 of SI) shows that application of ACPs leads to MAE reductions (%ΔMAE) in the range of about 27–67% for the NCI validation subsets containing complexes featuring a mix of common interactions found in large molecular systems (BlindNCI, DES15K, and NENCI-2021 data sets with a total of 17,413 data points). This range of %ΔMAE for the subsets used in the validation resembles the %ΔMAE obtained for data sets in training set with mixed character interactions such as S22x5, S66x8, and S66a8. Regarding hydrogen bonding interactions, the CE20, WaterOrg, and H2O20Bind10 subsets of the validation set feature these types of interactions. The MAEs for these three subsets are improved significantly on the application of ACPs by about 62–92%, depending on the method. The large reduction in error observed for mixed character and hydrogen bonding interactions in the validation are probably a consequence of the fact that the training set contains more data points of these two kinds than any other interaction type. In any case, the similarity between the error reduction in the validation and training set suggests that the proposed ACPs are fairly robust regarding these interactions, i.e., they can be applied to similar systems outside the training set.

Regarding the other interaction types, an assessment of the ACP performance in the validation stage for π-π stacking and pnicogen bonding interactions could not be carried out due to the scarcity of high-level reference data in the literature. This scarcity compelled us to include all available systems containing these interaction types in the training set instead of reserving them for validation. The ACPs were further validated on systems containing other interaction types such as halogen bonding, chalcogen bonding, and close contact repulsions (XB45, CHAL336, R160x6, and R739x5 data sets) that were not specifically part of the training set. The MAEs for the data sets representing halogen bonding were mostly deteriorated (by

>60%) on the application of ACPs, except for BLYP-D3/6-31G*, where the MAE was improved by 56%. Application of ACPs to subsets representing chalcogen bonding led to a decrease in the MAEs of uncorrected methods by 6–42%, most likely due to the presence of O and S containing complexes in the training set that were not purely chalcogen bonded. On the other hand, the MAEs of the uncorrected methods for the subsets representing close contact repulsions were initially low (0.50–0.97 kcal/mol), and the application of ACPs led to either increase (by 9–30%) or decrease (by 9–33%). All these findings suggest an under-representation of halogen bonding, chalcogen bonding, and close contact repulsions in the training set compared to other interaction types. Therefore, future ACP development work will require more such systems to be included in the training set in order to increase the diversity and robustness of the resulting ACPs.

The application of ACPs to complexes containing at least one monomer with negative charge in the validation set led to mostly a reduction in the MAEs of the uncorrected methods by 19–75%. Nevertheless, applying the ACPs could only bring down the MAEs to values that were greater than 2 kcal/mol. This indicates that the performance of ACPs for anionic interactions in the validation set convey the same message as in the training set: even though errors decrease on the application of ACPs, the 6-31G* basis set is inappropriate for modeling anionic systems.

Finally, we tested the proposed ACPs for their performance regarding non-covalent interaction energies in some more challenging complexes that are significantly different from those in the training set. In particular, we used the C60dimer set of non-covalent interaction energies between C_{60} dimeric complexes and the L7, S12L, S30L, and Ni2021 sets containing interaction energies between relatively large supramolecular systems. These data sets provide a more stringent test for ACPs than the other validation subsets because of the large size of the systems involved (the absolute reference energies range from 25 kcal/mol to 416 kcal/mol) as well as the multiplicity of and cooperativity between the non-covalent interactions present in these systems. The application of ACPs led to an overall reduction in the MAEs of the underlying methods for these sets by 11–66%, with only a few exceptions. For example, ACPs for CAM-B3LYP-D3/6-31G* reduce the MAEs of 4 out of 5 data sets by about 30–51%, indicating once again the robustness of the corresponding ACPs. The application of ACPs with CAM-B3LYP-D3/6-31G* led to an increase in the MAE of the C60dimer set from 1.78 kcal/mol to 3.97 kcal/mol. Although the error increases, this result also demonstrates that in the case of failure due to the systems studied being wildly different from those on which the ACP were trained, the results from the ACP-corrected methods are far from being catastrophic.

(ii) Conformational energies

Another molecular property included in the training set is molecular conformational energies ("CONF") with 11,161 data points. The purpose of these data is to inform ACPs about how molecular motion along rotatable bonds and torsional angles involving various effects (π-conjugation, steric interactions, intramolecular hydrogen-bonding, and electron repulsion) influence the molecular potential energy surfaces. The CONF systems include peptides, nucleotides, carbohydrates, alcohols, hydrocarbons, $(H_2O)_{11}$ clusters, and other molecules representing pharmaceuticals, catalysts, synthetic precursors, industrial chemicals, and organic compounds. Interestingly, the overall MAE of all uncorrected double-ζ basis set DFT methods for CONF in the training set is below 1 kcal/mol, likely due to error cancellation. Application of ACPs further reduces the MAEs by about 15–29% and brings them down to relatively low values: 0.70 (BLYP-D3/6-31G*-ACP), 0.43 (M062X/6-31G*-ACP), and 0.45 kcal/mol (CAM-B3LYP-D3/6-31G*-ACP).

In order to understand the performance of ACPs for CONF in the training set, we take a closer look at the results of CAM-B3LYP-D3/6-31G*-ACP (see Figure S1 of SI). These ACPs perform well for most subsets and are particularly suitable for peptides (PEPCONF-Dipeptide, TPCONF, P76, YMPJ), carbohydrates (SCONF, DSCONF, SacchCONF, CCONF), alcohols (BCONF), the melatonin molecule (MCONF), $(H_2O)_{11}$ clusters (Undecamer125), and a mix of various medium-sized organic molecules (37Conf8). The reduction in MAE for these subsets ranges between 24–84%, with most of them generally showing an improvement greater than 47%. The MAE reduction for other CONF subsets, like those containing organic molecules that are drug-like or have industrial relevance (DCONF and MolCONF) range between 13–18%, probably because of the already quite low MAEs of the uncorrected method (0.51 kcal/mol for DCONF and 0.41 kcal/mol for MolCONF).

Despite the general improvement in the CAM-B3LYP-D3/6-31G* method for conformational energies caused by ACPs, some outliers exist. The conformational energy data sets where the errors of ACP-corrected CAM-B3LYP-D3/6-31G* are higher than the uncorrected method are mainly found in the data sets for which the uncorrected method already has very small MAEs. For example, the hydrocarbon conformer subsets (ACONF and PentCONF) have MAEs of only 0.04–0.06 kcal/mol with CAM-B3LYP-D3/6-31G*. The MAEs of the ACP-corrected method for these data sets are higher than the uncorrected method but are still below 0.50 kcal/mol. Some other outlier subsets have MAEs for the uncorrected method in the range 0.25–0.58 kcal/mol, with the MAE from the ACP-corrected method rising to 0.42–0.74 kcal/mol.

Regarding the performance of ACPs for conformational energies outside the training set, we used 9,491 conformational energy data points in the validation set. The application of ACPs to the CONF data in the validation set ("CONF-val") shows overall good performance. The ACPs reduce the MAEs of M062X/6-31G* and CAM-B3LYP-D3/6-31G* by about 5–37% for two of the large data sets containing various conformers of organic molecules, viz. MAEs for Torsion30 of 0.32 and 0.36 kcal/mol and for ANI1ccxCONF of 1.72 and 1.68 kcal/mol, with M062X/6-31G*-ACP and CAMB3LYP-D3/6-31G*-ACP, respectively. ACPs applied to BLYP-D3/6-31G* also result in a reduction of MAEs (22–28%) that brings the MAEs down to 0.47 kcal/mol for Torsion30 and 2.76 kcal/mol for ANI1ccxCONF.

The performance of ACPs in the representative Torsion30 and ANI1ccxCONF validation examples shows that ACPs generally perform well for CONF data points that share similarities with those used in the training set. However, systems that are significantly different from the training set compilation result in somewhat higher but not catastrophic errors, for example in the data set containing conformational energies of peptide model systems with a disulfide-bridged bond where the ACPs fail to reduce the MAEs for the corresponding validation subset. Nevertheless, ACPs work well for other peptide conformational systems in the validation set, with MAE reductions in the range of 33–63% for CAM-B3LYP-D3/6-31G*-ACP in the case of conformational energies of other peptide systems (tripeptide, cyclic, and those that show bioactive functionality).

(iii) Reaction energies

We turn our attention now to covalent properties involving bond breaking and formation. One of such properties included in the training set are chemical reaction energies ("RE"). The training set contains 8,315 RE data points. Application of ACPs to the RE subsets in the training set shows an overall reduction in the MAEs of the double-ζ basis set DFT methods by about 29–34%. This improvement in the MAE mainly reflects the MAE decrease observed in two specific RE subsets, MN-RE (7,555 data points) and BH9-RE (449 data points). These subsets contain data points representing a variety of chemical reactions. For instance, the BH9-RE subset contains reaction energies of the nine most common elementary reactions encountered in organic and bioorganic chemistry. Besides MN-RE and BH9-RE, the ACPs also perform generally well (MAE reductions between 11–87% for M062X/6-31G*-ACP and CAM-B3LYP-D3/6-31G*-ACP) for other RE subsets of various types of reactions with fewer data points, with only a few exceptions. The decrease in the MAEs for reaction energies is also observed for the relatively large RE validation data set ("RE-val"), composed of a subset of the W4-17-RE set with 5,205 data points. For this data set, which was not used in the ACP training, the MAEs decrease by about 31–50% on application of

ACPs compared to the uncorrected double-ζ basis set DFT methods. This indicates that the ACPs are successful and robust for reaction energies.

(iv) Barrier heights

Besides reaction energies, the training set also included barrier heights of chemical reactions (“BH”) in order to make the eventual ACP-corrected methods usable for kinetic studies. The total number of BH data points contributes nearly 50% of the training set data (58,197) and is the most dominant property overall. The application of ACPs reduces the overall MAE of the parent methods for barrier heights by about 27–50%. This improvement originates mostly from the good performance of ACPs (MAE reductions by 13–52%) on the four main BH data sets: Grambow2020-B97D3 (32,722 data points), Grambow2020-ωB97XD3 (23,922 data points), BH9 (898), and E2SN2 (418). Note that the BH9 data set was designed recently[74] to be used for the particular purpose of developing the ACPs in this work and contains various model reactions that increase the diversity of the BH data in the training set.

For the other BH data sets in the training set besides BH9, E2SN2, and the two Grambow2020 subsets, the performance of ACPs is also quite good. ACPs mostly bring down the MAEs for the other subsets by about 22–94%. Contrary to most other BH data points in the training set, the INV24 and BHROT27 subsets feature barriers for processes that do not involve bond breaking or formation. The application of ACPs on these two data sets do not show a significant reduction in MAE compared to the uncorrected methods.

To validate the performance of ACPs on barrier height prediction, we applied the uncorrected methods and their ACP-corrected counterparts to a data set with 88 data points (WaterOrgBH) not included in the training set. This data set contains to-be-published barrier height reference data for pericyclic-type reactions in the absence and presence of water clusters, calculated at the DLPNO-CCSD(T)/CBS level of theory. The MAEs of double-ζ basis set DFT methods after application of ACPs to the WaterOrgBH data set decrease from 14.46 to 9.21 kcal/mol (BLYP-D3/6-31G*), 2.31 to 1.56 kcal/mol (M062X/6-31G*), and 1.70 to 1.66 kcal/mol (CAM-B3LYP-D3/6-31G*), indicating that ACPs are likely to perform well when they are applied to the calculation of barrier heights outside the training set.

As an additional test, we also applied the proposed ACPs to the systems in the work of Bistoni *et al.*[75], involving the Baeyer–Villiger reaction catalyzed by the cyclohexanone monooxygenase enzyme. The geometries used for ACP testing are the relevant stationary points of the reaction (reactant, intermediate, transition state, and product), obtained independently with three active site models of increasing size. The reference data for the relative energies along the reaction profile were obtained with DLPNO-CCSD(T0)/def2-TZVPP. Point charges were used to model the electrostatic potential from the surrounding protein environment. The errors in the calculated reaction energies and barrier heights with and without ACPs for the three differently sized active sites are shown in Table 3. The errors of various methods relative to DLPNO-CCSD(T0)/def2-TZVPP data in Table 3 demonstrates that the ACP-corrected methods yield lower errors in the predicted relative energies than the uncorrected methods for most stationary points along the reaction profile of different active site sizes, with only a few exceptions. Upon application of ACPs for the barrier height prediction of the smallest active site model (99 atoms), the errors in the barrier heights (energy of transition state relative to reactant) drop from -22.53 to -7.22 kcal/mol (BLYP-D3/6-31G*), -4.74 to 1.67 kcal/mol (M062X/6-31G*), and -5.14 to -1.09 kcal/mol (CAM-B3LYP-D3/6-31G*).

Table 3. Reference data (calculated using DLPNO-CCSD(T0)/def2-TZVPP) and errors relative to the reference data yielded by various methods (uncorrected and ACP-corrected double-ζ basis set DFT methods) for the relative energies along the reaction profile of Baeyer–Villiger reaction catalyzed by the cyclohexanone monooxygenase enzyme.[a,b,c]

Method	QM region with 99 atoms			QM region with 206 atoms			QM region with 307 atoms		
	Intermediate	Transition state	Product	Intermediate	Transition state	Product	Intermediate	Transition state	Product
DLPNO-CCSD(T0)/ def2-TZVPP	**-4.50**	**8.10**	**-68.60**	**-2.00**	**10.50**	**-71.10**	**-1.50**	**9.90**	**-71.90**
BLYP-D3/6-31G*	-9.4	-22.53	0.25	-2.98	-16.2	1.56	-1.08	-15.95	0.43
BLYP-D3/6-31G*-ACP	-1.4	-7.22	-1.94	0.94	-3.6	-1.67	2.72	-3.41	-2.22
M062X/6-31G*	-11.45	-4.74	-5.64	-8.5	0.03	-5.08	-6.22	1.06	-5.81
M062X/6-31G*-ACP	-4.36	1.67	1.94	-2.96	4.97	2.54	-0.41	5.82	2.42
CAM-B3LYP-D3/6-31G*	-8.36	-5.14	-3.65	-3.61	0.47	-2.35	-1.49	1.2	-3.24
CAM-B3LYP-D3/6-31G*-ACP	-2.71	-1.09	0.92	-0.63	2.67	1.73	1.85	3.3	1.56

[a] all energies were calculated in the presence of point charges with kcal/mol units, [b] the energies along the reaction profile were calculated relative to the reactant, [c] the geometries and point charges were taken from the work of Bistoni *et al.*[75]

(v) Bond separation energies and other properties

Lastly, we analyze ACP performance for bond separation energies ("BSE"). In our training set, we included 4,502 bond separation energies from our recently developed BSE49[76] data set. BSE49 contains the reaction energies associated with the formation of radical species upon homolytic cleavage of 49 unique single bonds with various functional group substitutions. The BSE49 set was designed so that the 49 bonds in question represent all single bonds between unique combinations of the ten atoms for which ACPs are being developed in this work. The lack of bond separation energy data in the literature was the main motivation behind the creation of the BSE49 data set. Therefore, further validation of ACPs on data outside the training set was not possible. Application of our ACPs leads to a reduction in the MAEs compared to the parent methods in BSE49 by about 26–43%, suggesting that ACPs offer an efficient way of modeling bond separation reactions.

Other chemical properties that were included in the ACP training set were: molecular isomerization energies ("ISOM"), total atomization energies ("TAE"), and molecular deformation energies ("DEF"). A few ISOM and TAE data sets from the literature were included in the training set, mainly to ensure that ACPs do not lead to a degradation in the functional performance for these properties. The results from Figure S1 in the SI show that the change in MAE for each of these properties upon application of the ACPs are reasonably acceptable. For example, the overall MAEs of ISOM and TAE with CAM-B3LYP-D3/6-31G* are respectively reduced by about 33% and 28%. Similarly, with M06-2X/6-31G* the MAEs are reduced by about 34% for ISOM and 8% for TAE. For BLYP-D3/6-31G*, the MAE is only reduced for ISOM by 42% but an increase in observed for the MAE of TAE by only 10%.

In addition, a large set of DEF data points were also included in the training set. These data points represent energy differences between a molecule at its equilibrium geometry and the same molecule deformed along its various normal modes. Our intention when we included these data in the training set was to improve the description of molecular potential energy surfaces to obtain reasonably accurate geometries and energy derivatives or, at least, prevent the ACP fitting procedure from deteriorating the performance of the uncorrected method for these properties. In this connection, we performed transition state searches and geometry optimizations on the stationary point structures (reactants, products, and transition states) of 100 representative chemical reactions taken from the BH9 data set. Application of ACPs resulted in geometries with average root-mean-square-deviations (RMSDs) ranging between 0.110–0.131 Å (Table S6 of SI) compared to the reference geometries. These average RMSDs are only slightly higher than the uncorrected double-ζ basis set DFT methods (0.068–0.123 Å). Note that the geometries yielded by DFT-D3 methods with a double-ζ basis set have been shown in the literature to be reasonably close to that obtained with quadruple-ζ basis sets.[22]

(vi) Overall ACP performance

In this subsection, we briefly discuss the overall performance of ACPs. First, we consider the various biases of the base method as measured by the mean signed errors (MSE) and standard deviations (SD) shown in Figure 1. Regarding non-covalent interaction energies, the figure shows that the uncorrected methods tend to over-estimate interaction energies leading to negative MSEs. Application of ACPs corrects for this bias and reduces both the MSE and the error spread (SD). For conformational energies, the MSEs and SDs of the base methods are reasonably low, and application of ACPs results in a very small change for these quantities. For conformational energies, ACPs bring down MSEs and SDs. Covalent properties such as reaction energies, barrier heights, and bond separation energies are mostly under-estimated (positive MSEs) with a substantial error spread. In the particular case of barrier heights and the BLYP functional, this may be attributed to delocalization error[74], which is partially remedied by the application of ACPs. For the covalent properties, application of the ACPs also decreases both the MSE and SDs substantially.

Lastly, we compare the performance of uncorrected and ACP-corrected double-ζ basis set DFT methods with the same DFT methods using large basis sets. We present the MAEs of the various methods in Table 4. Only a few representative data sets for which nearly complete basis set DFT results have been reported in the literature are shown in the table. For non-covalent properties (interaction and conformational energies), the MAEs of ACP-corrected methods are slightly higher than the same methods at complete basis set limit. However, a particular exception is seen in the case of large hydrogen-bonded water cluster sets, Water38 and H2O20Bind10 with CAM-B3LYP-D3/6-31G*-ACP where the MAEs are lower than CAM-B3LYP-D3/def2-QZVPDD by almost a factor of two. In the case of covalent properties (reaction energies and barrier heights), the MAEs of M062X/6-31G*-ACP and CAM-B3LYP-D3/6-31G*-ACP methods are lower than the same functionals with def2-QZVPDD basis set for a few cases like BH9-RE, BSR36, and PX13. The MAEs with CAM-B3LYP-D3/6-31G*-ACP is also lower than CAM-B3LYP-D3/def2-QZVPDD for two other data sets, namely HTBH38 and WCPT27. Another interesting result is observed for the ACP-corrected BLYP-D3/6-31G* method, where the MAEs of data sets containing reaction energies and barrier heights, with only a few exceptions, are generally lower than BLYP-D3/def2-QZVPDD MAEs, indicating that ACPs perform well in mitigating errors other than basis set incompleteness like those arising from the approximations in the density functionals itself. The overall comparison shown in Table 4 demonstrates that ACP-corrected methods have a performance close to nearly complete basis set DFT or sometimes even better, but at a ca. one order of magnitude lower computational time.

Table 4. Comparison of the mean absolute errors (MAEs) in kcal/mol of various 6-31G*, 6-31G*-ACP, and nearly complete basis set DFT methods for selected data sets.

Data set[a]	BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	BLYP-D3/def2-QZVPDD[b]	M062X/6-31G*	M062X/6-31G*-ACP	M062X/def2-QZVPDD[b]	CAM-B3LYP-D3/6-31G*	CAM-B3LYP-D3/6-31G*-ACP	CAM-B3LYP-D3/def2-QZVPDD[b]
Non-covalent interaction energies:									
S66x8[86–88] (528)	1.67	0.60	0.15	0.94	0.52	0.30	1.56	0.52	0.29
Sulfurx8[98] (104)	1.15	0.33	0.19	0.60	0.34	0.20	0.99	0.36	0.16
3B-69-DIM[96] (207)	1.65	0.61	0.24	0.76	0.48	0.43	1.68	0.35	0.31
BzDC215[102] (170)	0.94	0.44	0.14	0.63	0.22	0.22	0.82	0.33	0.15
Water38[124] (38)	36.41	1.39	1.30	26.22	4.81	1.62	36.38	3.39	5.87
CE20[189,190] (20)	21.46	2.86	1.02	15.18	3.30	1.16	21.40	3.81	3.27
H2O20Bind10[200] (10)	145.88	24.40	4.50	107.24	15.19	3.35	154.83	13.15	22.12
Molecular conformational energies:									
YMPJ[143] (495)	1.08	0.85	0.53	0.72	0.62	0.38	1.06	0.52	0.34
BCONF[152] (64)	2.55	0.85	0.34	1.75	0.45	0.12	2.44	0.39	0.34
SCONF[14,147] (17)	4.03	0.98	0.41	2.47	0.63	0.18	3.29	0.75	0.19
PentCONF[153] (342)	0.30	0.21	0.36	0.13	0.17	0.11	0.06	0.42	0.08
Reaction energies:									
BH9-RE[74] (449)	5.39	3.12	7.15[c]	3.40	2.02	2.76[d]	3.49	2.43	3.14[e]
BH76RC[162–164] (30)	8.72	4.02	3.28	7.35	3.98	0.86	8.44	4.43	1.61
NBPRC[161,164,181] (6)	2.81	1.77	3.08	2.13	3.34	1.10	1.73	1.84	2.05
BSR36[160,161] (36)	2.74	2.61	3.33	3.62	0.77	3.51	4.81	1.03	4.26
WCPT6[180] (6)	3.71	1.37	1.04	2.74	2.42	0.80	3.42	2.22	0.89
CR20[166] (20)	2.46	2.70	9.63	3.92	2.61	1.75	4.52	3.34	2.52
DIE60[157] (60)	2.39	1.12	1.43	1.16	0.67	0.57	1.29	1.05	0.57
Barrier heights:									
BH9[74] (898)	12.72	7.40	8.66[c]	3.43	2.98	2.27[d]	4.16	2.88	2.37[e]
BHPERI26[164,184] (26)	7.24	4.47	3.58	7.24	4.47	1.35	2.25	2.60	2.37
CRBH20[188] (20)	13.52	1.34	16.56	13.52	1.34	1.32	2.31	1.61	1.18
DBH24[185,186] (24)	10.99	5.23	8.25	10.99	5.23	0.85	6.00	4.45	2.66
HTBH38[163] (38)	10.01	5.09	8.79	10.01	5.09	1.08	4.53	2.58	3.61
NHTBH38[162] (38)	13.59	5.54	8.95	13.59	5.54	1.29	8.14	5.37	2.71
PX13[189,190] (13)	33.76	2.17	8.88	33.76	2.17	6.11	29.49	3.59	8.06
WCPT27[180] (27)	12.81	3.18	6.53	12.81	3.18	2.92	9.66	2.49	3.43

[a)] details about the data sets can be found in Table S1 of the Supporting Information, [b)] from Reference 13, [c)] represents the value obtained with BLYP-XDM/def2-QZVPP in Reference 74, [d)] represents the value obtained with M062X/def2-QZVPP in Reference 74, [e)] represents the value obtained with CAM-B3LYP-XDM/def2-QZVPP in Reference 74.

4. Conclusions

The use of very accurate quantum mechanical methods, such as nearly complete basis set wavefunction theory methods, is not practical for applications involving large systems because of the steep scaling of their computational cost with system size. A low-cost alternative is the use of small basis set density-functional theory (DFT) methods, but this requires that the inherent shortcomings in these methods be mitigated. The focus of this article is to improve the performance of small basis set DFT methods without sacrificing their computational efficiency.

We developed and applied atom-centered potentials (ACPs) to mitigate the shortcomings of BLYP-D3, M062X, and CAM-B3LYP-D3 methods, in combination with the 6-31G* basis set. We expect these shortcomings to be primarily the error from the choice of density functional approximation and basis set incompleteness from the limited size of the basis set.

The ACPs presented in this work were developed for ten elements (H, B, C, N, O, F, Si, P, S, Cl). The parametrization of the ACPs was carried out using a regularized linear least-squares fitting procedure (the LASSO regression method) using a training set of 118,655 data points calculated mostly using wavefunction theory methods extrapolated to the complete basis set limit. The main molecular properties in the training set were non-covalent interaction energies, molecular conformational energies, reaction energies, barrier heights, and bond separation energies. The performance of the proposed ACPs was assessed using the training set and an additional validation set containing 42,567 data points not used during the ACP training.

Our assessment of the new ACP-corrected methods suggests that the ACPs reduce the mean absolute errors (MAEs) of double-ζ basis set DFT methods for most subsets, and in general, lead to an improved description of all molecular properties in the training set, indicating that ACPs successfully mitigate the deficiencies of the parent double-ζ basis set DFT methods. The best performing ACP-based method, i.e., CAM-B3LYP-D3/6-31G*-ACP, yields mean absolute errors, relative to high-level of theory, of around 0.7 kcal/mol for non-covalent interaction energies, 0.4 kcal/mol for conformational energies, 5.6 kcal/mol for reaction energies, 3.0 kcal/mol for barrier heights, and 2.5 kcal/mol for bond separation energies.

Further analysis of the performance of ACP-corrected methods on the validation set shows that ACPs are relatively robust, i.e., they are suitable for applications in systems outside the training set. However, there is a performance penalty when the system under consideration is significantly different from our training set, although no catastrophic results were obtained at any point. This observation suggests that increasing the size of the training set and incorporating more diversity is a straightforward way to improve the ACPs. In addition, using basis sets containing diffuse functions could be a way to overcome the limitations of the present methods in the description of negatively charged systems. The proposed new

ACPs correct small basis set DFT methods and improve the accuracy of their parent methods, and allow carrying out quantum mechanical calculations of large systems at a reasonably low computational cost.

References

(1) Burke, K. Perspective on Density Functional Theory. *J. Chem. Phys.* **2012**, *136* (15), 150901.

(2) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112* (1), 289–320.

(3) DiLabio, G. A.; Otero-de-la-Roza, A. Noncovalent Interactions in Density Functional Theory; In *Reviews in Computational Chemistry*; John Wiley & Sons Inc., 2016; pp 1–97.

(4) Riley, K. E.; Hobza, P. Noncovalent Interactions in Biochemistry. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 3–17.

(5) Goerigk, L. A Comprehensive Overview of the DFT-D3 London-Dispersion Correction. In *Non-Covalent Interactions in Quantum Chemistry and Physics: Theory and Applications*; Elsevier Inc., **2017**, pp 195–219.

(6) Johnson, E. R. The Exchange-Hole Dipole Moment Dispersion Model. In *Non-Covalent Interactions in Quantum Chemistry and Physics: Theory and Applications*; Elsevier Inc., 2017; pp 169–194.

(7) Hermann, J.; DiStasio Jr., R. A.; Tkatchenko, A. First-Principles Models for van Der Waals Interactions in Molecules and Materials: Concepts, Theory, and Applications. *Chem. Rev.* **2017**, *117* (6), 4714–4758.

(8) Stöhr, M.; Van Voorhis, T.; Tkatchenko, A. Theory and Practice of Modeling van Der Waals Interactions in Electronic-Structure Calculations. *Chem. Soc. Rev.* **2019**, *48*, 4118–4154.

(9) Ambrosetti, A.; Alfè, D.; DiStasio Jr., R. A..; Tkatchenko, A. Hard Numbers for Large Molecules: Toward Exact Energetics for Supramolecular Systems. **2014**, *5* (5), 849–855.

(10) Otero-de-la-Roza, A.; Johnson, E. R. Predicting Energetics of Supramolecular Systems Using the XDM Dispersion Model. *J. Chem. Theory Comput.* **2015**, *11* (9), 4033–4040.

(11) Risthaus, T.; Grimme, S. Benchmarking of London Dispersion-Accounting Density Functional Theory Methods on Very Large Molecular Complexes. *J. Chem. Theory Comput.* **2013**, *9* (3), 1580–1591.

(12) Antony, J.; Sure, R.; Grimme, S. Using Dispersion-Corrected Density Functional Theory to Understand Supramolecular Binding Thermodynamics. *Chem. Commun.* **2015**, *51* (10), 1764–1774.

(13) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115* (19), 2315–2372.

(14) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A Look at the Density Functional Theory Zoo with the Advanced GMTKN55 Database for General Main Group Thermochemistry, Kinetics and Noncovalent Interactions. *Phys. Chem. Chem. Phys.* **2017**, *19* (48), 32184–32215.

(15) Cui, Q. Perspective: Quantum Mechanical Methods in Biochemistry and Biophysics. *J. Chem. Phys.* **2016**, *145* (14), 140901.

(16) Ratcliff, L. E.; Mohr, S.; Huhs, G.; Deutsch, T.; Masella, M.; Genovese, L. Challenges in Large Scale Quantum Mechanical Calculations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *7* (1), e1290.

(17) Brandenburg, J. G.; Burke, K.; Civalleri, B.; Cole, D. J.; Csányi, G.; David, G.; Gidopoulos, N. I.; Gowland, D.; Helgaker, T.; Herbst, M. F.; Hourahine, B.; Irons, T. J. P; Jacob, C. R.; Loos, P.-F.; Mehta, N.; Mulay, M. R.; Neugebauer, J.; Pernal, K.; Pribram-Jones, A.; Romaniello, P.; Ryder, M. R.; Savin, A.; Sirbu, D.; Skylaris, C.-K.; Truhlar, D. G.; Wetherell, J.; Yang, W. Challenges for Large Scale Simulation: General Discussion. *Faraday Discuss.* **2020**, *224* (0), 309–332.

(18) Sherrill, C. D. Frontiers in Electronic Structure Theory. *J. Chem. Phys.* **2010**, *132* (11), 110902.

(19) Hofer, T. S. From Macromolecules to Electrons—Grand Challenges in Theoretical and Computational Chemistry. *Front. Chem.* **2013**, *1*, 6.

(20) Grimme, S.; Schreiner, P. R. Computational Chemistry: The Fate of Current Methods and Future Challenges. *Angew. Chemie Int. Ed.* **2018**, *57* (16), 4170–4176.

(21) Houk, K. N.; Liu, F. Holy Grails for Computational Organic Chemistry and Biochemistry. *Acc. Chem. Res.* **2017**, *50* (3), 539–543.

(22) Hostaš, J.; Řezáč, J. Accurate DFT-D3 Calculations in a Small Basis Set. *J. Chem. Theory Comput.* **2017**, *13* (8), 3575–

3585.

(23) Sure, R.; Brandenburg, J. G.; Grimme, S. Small Atomic Orbital Basis Set First-Principles Quantum Chemical Methods for Large Molecular and Periodic Systems: A Critical Analysis of Error Sources. *ChemistryOpen* **2016**, *5* (2), 94–109.

(24) Kruse, H.; Goerigk, L.; Grimme, S. Why the Standard B3LYP/6-31G* Model Chemistry Should Not Be Used in DFT Calculations of Molecular Thermochemistry: Understanding and Correcting the Problem. *J. Org. Chem.* **2012**, *77* (23), 10824–10834.

(25) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate *Ab Initio* Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104.

(26) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1456–1465.

(27) Kruse, H.; Grimme, S. A Geometrical Correction for the Inter- and Intra-Molecular Basis Set Superposition Error in Hartree-Fock and Density Functional Theory Calculations for Large Systems. *J. Chem. Phys.* **2012**, *136* (15), 154101.

(28) Sure, R.; Grimme, S. Corrected Small Basis Set Hartree-Fock Method for Large Systems. *J. Comput. Chem.* **2013**, *34* (19), 1672–1685.

(29) Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. Consistent Structures and Interactions by Density Functional Theory with Small Atomic Orbital Basis Sets. *J. Chem. Phys.* **2015**, *143* (5), 054107.

(30) Brandenburg, J. G.; Caldeweyher, E.; Grimme, S. Screened Exchange Hybrid Density Functional for Accurate and Efficient Structures and Interaction Energies. *Phys. Chem. Chem. Phys.* **2016**, *18* (23), 15519–15523.

(31) Pracht, P.; Grant, D. F.; Grimme, S. Comprehensive Assessment of GFN Tight-Binding and Composite Density Functional Theory Methods for Calculating Gas-Phase Infrared Spectra. *J. Chem. Theory Comput.* **2020**, *16* (11), 7044–7060.

(32) Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. B97-3c: A Revised Low-Cost Variant of the B97-D Density Functional Method. *J. Chem. Phys.* **2018**, *148* (6), 064104.

(33) Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. r2SCAN-3c: A “Swiss Army Knife” Composite Electronic-Structure Method . *J. Chem. Phys.* **2021**, *154* (6), 064103.

(34) Kulik, H. J.; Seelam, N.; Mar, B. D.; Martínez, T. J. Adapting DFT+ *U* for the Chemically Motivated Correction of Minimal Basis Set Incompleteness. *J. Phys. Chem. A* **2016**, *120* (29), 5939–5949.

(35) Schneebeli, S. T.; Bochevarov, A. D.; Friesner, R. A. Parameterization of a B3LYP Specific Correction for Noncovalent Interactions and Basis Set Superposition Error on a Gigantic Data Set of CCSD(T) Quality Noncovalent Interaction Energies. *J. Chem. Theory Comput.* **2011**, *7* (3), 658–668.

(36) García, J. S.; Brémond, É.; Campetella, M.; Ciofini, I.; Adamo, C. Small Basis Set Allowing the Recovery of Dispersion Interactions with Double-Hybrid Functionals. *J. Chem. Theory Comput.* **2019**, *15* (5), 2944–2953.

(37) Li, W.; Miao, W.; Cui, J.; Fang, C.; Su, S.; Li, H.; Hu, L.; Lu, Y.; Chen, G. Efficient Corrections for DFT Noncovalent Interactions Based on Ensemble Learning Models. *J. Chem. Inf. Model.* **2019**, *59* (5), 1849–1857.

(38) Mehta, N.; Goerigk, L.; Mehta, N.; Goerigk, L. Assessing the Applicability of the Geometric Counterpoise Correction in B2PLYP/Double-ζ Calculations for Thermochemistry, Kinetics, and Noncovalent Interactions. *Aust. J. Chem.* **2021**. https://doi.org/10.1071/CH21133

(39) Tirri, B.; Ciofini, I.; Sancho-García, J. C.; Adamo, C.; Brémond, É. Computation of Covalent and Noncovalent Structural Parameters at Low Computational Cost: Efficiency of the DH-SVPD Method. *Int. J. Quantum Chem.* **2020**, *120* (13), e26233.

(40) Witte, J.; Neaton, J. B.; Head-Gordon, M. Effective Empirical Corrections for Basis Set Superposition Error in the Def2-SVPD Basis: gCP and DFT-C. *J. Chem. Phys.* **2017**, *146* (23), 234105.

(41) DiLabio, G. A. Atom-centered Potentials for Noncovalent Interactions and Other Applications. In *Non-Covalent Interactions in Quantum Chemistry and Physics: Theory and Applications*; Elsevier Inc., **2017**, pp 221–240.

(42) Prasad, V. K.; Otero-de-la-Roza, A.; DiLabio, G. A. Atom-Centered Potentials with Dispersion-Corrected Minimal-Basis-Set Hartree–Fock: An Efficient and Accurate Computational Approach for Large Molecular Systems. *J. Chem. Theory Comput.* **2018**, *14* (2), 726–738.

(43) Otero-de-la-Roza, A.; Dilabio, G. A. Improved Basis-Set Incompleteness Potentials for Accurate Density-Functional Theory Calculations in Large Systems. *J. Chem. Theory Comput.* **2020**, *16* (7), 4176–4191.

(44) Otero-de-la-Roza, A.; DiLabio, G. A. Transferable Atom-Centered Potentials for the Correction of Basis Set Incompleteness Errors in Density-Functional Theory. *J. Chem. Theory Comput.* **2017**, *13* (8), 3505–3524.

(45) van Santen, J. A.; DiLabio, G. A. Dispersion Corrections Improve the Accuracy of Both Noncovalent and Covalent

Interactions Energies Predicted by a Density-Functional Theory Approximation. *J. Phys. Chem. A* **2015**, *119* (25), 6703–6713.

(46) DiLabio, G. A.; Koleini, M. Dispersion-Correcting Potentials Can Significantly Improve the Bond Dissociation Enthalpies and Noncovalent Binding Energies Predicted by Density-Functional Theory. *J. Chem. Phys.* **2014**, *140* (18), 18A542.

(47) DiLabio, G. A.; Koleini, M.; Torres, E. Extension of the B3LYP–Dispersion-Correcting Potential Approach to the Accurate Treatment of Both Inter- and Intra-Molecular Interactions. *Theor. Chem. Acc.* **2013**, *132* (10), 1389.

(48) Torres, E.; DiLabio, G. A. A (Nearly) Universally Applicable Method for Modeling Noncovalent Interactions Using B3LYP. *J. Phys. Chem. Lett.* **2012**, *3* (13), 1738–1744.

(49) Mackie, I. D.; DiLabio, G. A. Interactions in Large, Polyaromatic Hydrocarbon Dimers: Application of Density Functional Theory with Dispersion Corrections. *J. Phys. Chem. A* **2008**, *112* (43), 10968–10976.

(50) DiLabio, G. A. Accurate Treatment of van Der Waals Interactions Using Standard Density Functional Theory Methods with Effective Core-Type Potentials: Application to Carbon-Containing Dimers. *Chem. Phys. Lett.* **2008**, *455* (4–6), 348–353.

(51) Mackie, I. D.; DiLabio, G. A. Accurate Dispersion Interactions from Standard Density-Functional Theory Methods with Small Basis Sets. *Phys. Chem. Chem. Phys.* **2010**, *12* (23), 6092.

(52) Torres, E.; DiLabio, G. A. Density-Functional Theory with Dispersion-Correcting Potentials for Methane: Bridging the Efficiency and Accuracy Gap between High-Level Wave Function and Classical Molecular Mechanics Methods. *J. Chem. Theory Comput.* **2013**, *9* (8), 3342–3349.

(53) Mackie, I. D.; DiLabio, G. A. CO $_2$ Adsorption by Nitrogen-Doped Carbon Nanotubes Predicted by Density-Functional Theory with Dispersion-Correcting Potentials. *Phys. Chem. Chem. Phys.* **2011**, *13* (7), 2780–2787.

(54) Holmes, J. D.; Otero-de-la-Roza, A.; DiLabio, G. A. Accurate Modeling of Water Clusters with Density-Functional Theory Using Atom-Centered Potentials. *J. Chem. Theory Comput.* **2017**, *13* (9), 4205–4215.

(55) Prasad, V. K.; Otero-de-la-Roza, A.; DiLabio, G. A. Performance of Small Basis Set Hartree–Fock Methods for Modeling Non-Covalent Interactions. *Electron. Struct.* **2021**, *3* (3), 034007.

(56) Cao, X.; Dolg, M. Pseudopotentials and Modelpotentials. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (2), 200–210.

(57) Dolg, M.; Cao, X. Relativistic Pseudopotentials: Their Development and Scope of Applications. *Chem. Rev.* **2012**, *112* (1), 403–480.

(58) Prasad, V. K.; Otero-de-la-Roza, A.; DiLabio, G. A. Fast and Accurate Quantum Mechanical Modeling of Large Molecular Systems Using Small Basis Set Hartree–Fock Corrected with Atom-Centered Potentials. **2021**, in preparation.

(59) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-consistent Molecular Orbital Methods. XXIII. A Polarization-type Basis Set for Second-row Elements. *J. Chem. Phys.* **1982**, *77* (7), 3654–3665.

(60) Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta* **1973**, *28* (3), 213–222.

(61) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100.

(62) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37* (2), 785–789.

(63) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120* (1–3), 215–241.

(64) Yanai, T.; Tew, D. P.; Handy, N. C. A New Hybrid Exchange–Correlation Functional Using the Coulomb-Attenuating Method (CAM-B3LYP). *Chem. Phys. Lett.* **2004**, *393* (1–3), 51–57.

(65) Density Functional Theory Poll. https://www.marcelswart.eu/dft-poll/ (accessed 2021-10-31)

(66) Goerigk, L.; Mehta, N. A Trip to the Density Functional Theory Zoo: Warnings and Recommendations for the User. *Aust. J. Chem.* **2019**, *72* (8), 563–573.

(67) Witte, J.; Neaton, J. B.; Head-Gordon, M. Push It to the Limit: Characterizing the Convergence of Common Sequences of Basis Sets for Intermolecular Interactions as Described by Density Functional Theory. *J. Chem. Phys.* **2016**, *144* (19), 194306.

(68) Tibshirani, R. Regression Shrinkage and Selection via the Lasso: A Retrospective. *J. R. Stat. Soc. Ser. B Stat. Methodol.*

2011, *73* (3), 273–282.

(69) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58* (1), 267–288.

(70) Osborne, M. R.; Presnell, B.; Turlach, B. A. On the LASSO and Its Dual. *J. Comput. Graph. Stat.* **2000**, *9* (2), 319–337.

(71) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; calmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 16, Revision B.01; Gaussian Inc.: Wallingford, CT, 2016.

(72) Johnson, E. R.; Otero-de-la-Roza, A.; Dale, S. G.; Dilabio, G. A. Efficient Basis Sets for Non-Covalent Interactions in XDM-Corrected Density-Functional Theory. *J. Chem. Phys.* **2013**, *139* (21), 214109.

(73) Papajak, E.; Truhlar, D. G. Efficient Diffuse Basis Sets for Density Functional Theory. *J. Chem. Theory Comput.* **2010**, *6* (3), 597–601.

(74) Prasad, V. K.; Pei, Z.; Edelmann, S.; Otero-de-la-Roza, A.; DiLabio, G. BH9, a New Comprehensive Benchmark Dataset for Barrier Heights and Reaction Energies: Assessment of Density Functional Approximations and Basis Set Incompleteness Potentials. *ChemRxiv* **2021**. https://doi.org/10.33774/CHEMRXIV-2021-MDJWX.

(75) Bistoni, G.; Polyak, I.; Sparta, M.; Thiel, W.; Neese, F. Toward Accurate QM/MM Reaction Barriers with Large QM Regions Using Domain Based Pair Natural Orbital Coupled Cluster Theory. *J. Chem. Theory Comput.* **2018**, *14* (7), 3524–3531.

(76) Prasad, V. K., Khalilian, H., Otero-de-la-Roza, A. & DiLabio, G. A. BSE49, a Diverse, High-Quality Benchmark Dataset of Separation Energies of Chemical Bonds. *Sci. Data* **2021**, in review.

(77) Thanthiriwatte, K. S.; Hohenstein, E. G.; Burns, L. A.; Sherrill, C. D. Assessment of the Performance of DFT and DFT-D Methods for Describing Distance Dependence of Hydrogen-Bonded Interactions. *J. Chem. Theory Comput.* **2011**, *7* (1), 88–96.

(78) Marshall, M. S.; Burns, L. A.; Sherrill, C. D. Basis Set Convergence of the Coupled-Cluster Correction, P2CCSD(T): Best Practices for Benchmarking Non-Covalent Interactions and the Attendant Revision of the S22, NBC10, HBC6, and HSG Databases. *J. Chem. Phys.* **2011**, *135* (19), 194102.

(79) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes. *J. Chem. Theory Comput.* **2009**, *5* (7), 1749–1760.

(80) Miriyala, V. M.; Řezáč, J. Description of Non-Covalent Interactions in SCC-DFTB Methods. *J. Comput. Chem.* **2017**, *38* (10), 688–697.

(81) Řezáč, J.; Hobza, P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *J. Chem. Theory Comput.* **2012**, *8* (1), 141–151.

(82) Řezáč, J. Non-Covalent Interactions Atlas Benchmark Data Sets: Hydrogen Bonding. *J. Chem. Theory Comput.* **2020**, *16* (4), 2355–2368.

(83) Řezáč, J. Non-Covalent Interactions Atlas Benchmark Data Sets 2: Hydrogen Bonding in an Extended Chemical Space. *J. Chem. Theory Comput.* **2020**, *16* (10), 6305–6316.

(84) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. Benchmark Database of Accurate (MP2 and CCSD(T) Complete Basis Set Limit) Interaction Energies of Small Model Complexes, DNA Base Pairs, and Amino Acid Pairs. *Phys. Chem. Chem. Phys.* **2006**, *8* (17), 1985–1993.

(85) Gráfová, L.; Pitoňák, M.; Řezáč, J.; Hobza, P. Comparative Study of Selected Wave Function and Density Functional Methods for Noncovalent Interaction Energy Calculations Using the Extended S22 Data Set. *J. Chem. Theory Comput.* **2010**, *6* (8), 2365–2376.

(86) Řezáč, J.; Riley, K. E.; Hobza, P. S66: A Well-Balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J. Chem. Theory Comput.* **2011**, *7* (8), 2427–2438.

(87) Řezáč, J.; Riley, K. E.; Hobza, P. Extensions of the S66 Data Set: More Accurate Interaction Energies and Angular-Displaced Nonequilibrium Geometries. *J. Chem. Theory Comput.* **2011**, *7* (11), 3466–3470.

(88) Dilabio, G. A.; Johnson, E. R.; Otero-de-la-Roza, A. Performance of Conventional and Dispersion-Corrected Density-Functional Theory Methods for Hydrogen Bonding Interaction Energies. *Phys. Chem. Chem. Phys.* **2013**, *15* (31), 12821–12828.

(89) Řezáč, J.; Dubecký, M.; Jurečka, P.; Hobza, P. Extensions and Applications of the A24 Data Set of Accurate Interaction Energies. *Phys. Chem. Chem. Phys.* **2015**, *17* (29), 19268–19277.

(90) Witte, J.; Goldey, M.; Neaton, J. B.; Head-Gordon, M. Beyond Energies: Geometries of Nonbonded Molecular Complexes as Metrics for Assessing Electronic Structure Approaches. *J. Chem. Theory Comput.* **2015**, *11* (4), 1481–1492.

(91) Řezáč, J.; Hobza, P. Describing Noncovalent Interactions beyond the Common Approximations: How Accurate Is the "Gold Standard," CCSD(T) at the Complete Basis Set Limit? *J. Chem. Theory Comput.* **2013**, *9* (5), 2151–2155.

(92) Sherrill, C. D.; Takatani, T.; Hohenstein, E. G. An Assessment of Theoretical Methods for Nonbonded Interactions: Comparison to Complete Basis Set Limit Coupled-Cluster Potential Energy Curves for the Benzene Dimer, the Methane Dimer, Benzene-Methane, and Benzene-H2S. *J. Phys. Chem. A* **2009**, *113* (38), 10146–10159.

(93) Hohenstein, E. G.; Sherrill, C. D. Effects of Heteroatoms on Aromatic π-π Interactions: Benzene-Pyridine and Pyridine Dimer. *J. Phys. Chem. A* **2009**, *113* (5), 878–886.

(94) Takatani, T.; Sherrill, C. D. Performance of Spin-Component-Scaled Møller-Plesset Theory (SCS-MP2) for Potential Energy Curves of Noncovalent Interactions. *Phys. Chem. Chem. Phys.* **2007**, *9* (46), 6106–6114.

(95) Smith, D. G. A.; Burns, L. A.; Patkowski, K.; Sherrill, C. D. Revised Damping Parameters for the D3 Dispersion Correction to Density Functional Theory. *J. Phys. Chem. Lett.* **2016**, *7* (12), 2197–2203.

(96) Řezáč, J.; Huang, Y.; Hobza, P.; Beran, G. J. O. Benchmark Calculations of Three-Body Intermolecular Interactions and the Performance of Low-Cost Electronic Structure Methods. *J. Chem. Theory Comput.* **2015**, *11* (7), 3065–3079.

(97) Copeland, K. L.; Tschumper, G. S. Hydrocarbon/Water Interactions: Encouraging Energetics and Structures from Dft but Disconcerting Discrepancies for Hessian Indices. *J. Chem. Theory Comput.* **2012**, *8* (5), 1646–1656.

(98) Mintz, B. J.; Parks, J. M. Benchmark Interaction Energies for Biologically Relevant Noncovalent Complexes Containing Divalent Sulfur. *J. Phys. Chem. A* **2012**, *116* (3), 1086–1092.

(99) Sanders, J. M. Optimal π-Stacking Interaction Energies in Parallel-Displaced Aryl/Aryl Dimers Are Predicted by the Dimer Heavy Atom Count. *J. Phys. Chem. A* **2010**, *114* (34), 9205–9211.

(100) Parrish, R. M.; Sherrill, C. D. Quantum-Mechanical Evaluation of π-π Versus Substituent-π Interactions in π Stacking: Direct Evidence for the Wheeler-Houk Picture. *J. Am. Chem. Soc.* **2014**, *136* (50), 17386–17389.

(101) Steinmann, S. N.; Corminboeuf, C. Exploring the Limits of Density Functional Approximations for Interaction Energies of Molecular Precursors to Organic Electronics. *J. Chem. Theory Comput.* **2012**, *8* (11), 4305–4316.

(102) Crittenden, D. L. A Systematic CCSD(T)Study of Long-Range and Noncovalent Interactions between Benzene and a Series of First- and Second-Row Hydrides and Rare Gas Atoms. *J. Phys. Chem. A* **2009**, *113* (8), 1663–1669.

(103) Hill, J. G.; Legon, A. C. On the Directionality and Non-Linearity of Halogen and Hydrogen Bonds. *Phys. Chem. Chem. Phys.* **2015**, *17* (2), 858–867.

(104) Řezáč, J.; Riley, K. E.; Hobza, P. Benchmark Calculations of Noncovalent Interactions of Halogenated Molecules. *J. Chem. Theory Comput.* **2012**, *8* (11), 4285–4292.

(105) Setiawan, D.; Kraka, E.; Cremer, D. Strength of the Pnicogen Bond in Complexes Involving Group VA Elements N, P, and AS. *J. Phys. Chem. A* **2015**, *119* (9), 1642–1656.

(106) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M. Estimated MP2 and CCSD(T) Interaction Energies of n-Alkane Dimers at the Basis Set Limit: Comparison of the Methods of Helgaker et Al. and Feller. *J. Chem. Phys.* **2006**, *124* (11), 114304.

(107) Granatier, J.; Pitoňák, M.; Hobza, P. Accuracy of Several Wave Function and Density Functional Theory Methods for Description of Noncovalent Interaction of Saturated and Unsaturated Hydrocarbon Dimers. *J. Chem. Theory Comput.* **2012**, *8* (7), 2282–2292.

(108) Burns, L. A.; Faver, J. C.; Zheng, Z.; Marshall, M. S.; Smith, D. G. A.; Vanommeslaeghe, K.; MacKerell, A. D.; Merz, K. M.; Sherrill, C. D. The BioFragment Database (BFDb): An Open-Data Platform for Computational Chemistry Analysis of Noncovalent Interactions. *J. Chem. Phys.* **2017**, *147* (16), 161727.

(109) Jakubec, D.; Hostaš, J.; Laskowski, R. A.; Hobza, P.; Vondrášek, J. Large-Scale Quantitative Assessment of Binding Preferences in Protein-Nucleic Acid Complexes. *J. Chem. Theory Comput.* **2015**, *11* (4), 1939–1948.

(110) Hostaš, J.; Jakubec, D.; Laskowski, R. A.; Gnanasekaran, R.; Řezáč, J.; Vondrášek, J.; Hobza, P. Representative Amino Acid Side-Chain Interactions in Protein-DNA Complexes: A Comparison of Highly Accurate Correlated Ab Initio

Quantum Mechanical Calculations and Efficient Approaches for Applications to Large Systems. *J. Chem. Theory Comput.* **2015**, *11* (9), 4086–4092.

(111) Jakubec, D.; Laskowski, R. A.; Vondrasek, J. Sequence-Specific Recognition of DNA by Proteins: Binding Motifs Discovered Using a Novel Statistical/Computational Analysis. *PLoS One* **2016**, *11* (7), e0158704.

(112) Stasyuk, O. A.; Jakubec, D.; Vondrášek, J.; Hobza, P. Noncovalent Interactions in Specific Recognition Motifs of Protein-DNA Complexes. *J. Chem. Theory Comput.* **2017**, *13* (2), 877–885.

(113) Kozmon, S.; Matuška, R.; Spiwok, V.; Koča, J. Three-Dimensional Potential Energy Surface of Selected Carbohydrates' CH/π Dispersion Interactions Calculated by High-Level Quantum Mechanical Methods. *Chem. - A Eur. J.* **2011**, *17* (20), 5680–5690.

(114) Kozmon, S.; Matuška, R.; Spiwok, V.; Koča, J. Dispersion Interactions of Carbohydrates with Condensate Aromatic Moieties: Theoretical Study on the CH-π Interaction Additive Properties. *Phys. Chem. Chem. Phys.* **2011**, *13* (31), 14215–14222.

(115) Stanković, I. M.; Blagojević Filipović, J. P.; Zarić, S. D. Carbohydrate – Protein Aromatic Ring Interactions beyond CH/π Interactions: A Protein Data Bank Survey and Quantum Chemical Calculations. *Int. J. Biol. Macromol.* **2020**, *157*, 1–9.

(116) Kumari, M.; Sunoj, R. B.; Balaji, P. V. Conformational Mapping and Energetics of Saccharide-Aromatic Residue Interactions: Implications for the Discrimination of Anomers and Epimers and in Protein Engineering. *Org. Biomol. Chem.* **2012**, *10* (21), 4186–4200.

(117) Černý, J.; Schneider, B.; Biedermannová, L. WatAA: Atlas of Protein Hydration. Exploring Synergies between Data Mining and: Ab Initio Calculations. *Phys. Chem. Chem. Phys.* **2017**, *19* (26), 17094–17102.

(118) Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; Sherrill, C. D.; Merz, K. M. Formal Estimation of Errors in Computed Absolute Interaction Energies of Protein-Ligand Complexes. *J. Chem. Theory Comput.* **2011**, *7* (3), 790–797.

(119) Kříž, K.; Řezáč, J. Benchmarking of Semiempirical Quantum-Mechanical Methods on Systems Relevant to Computer-Aided Drug Design. *J. Chem. Inf. Model.* **2020**, *60* (3), 1453–1460.

(120) Kruse, H.; Banáš, P.; Šponer, J. Investigations of Stacked DNA Base-Pair Steps: Highly Accurate Stacking Interaction Energies, Energy Decomposition, and Many-Body Stacking Effects. *J. Chem. Theory Comput.* **2019**, *15* (1), 95–115.

(121) Parker, T. M.; Sherrill, C. D. Assessment of Empirical Models versus High-Accuracy Ab Initio Methods for Nucleobase Stacking: Evaluating the Importance of Charge Penetration. *J. Chem. Theory Comput.* **2015**, *11* (9), 4197–4204.

(122) Banáš, P.; Mládek, A.; Otyepka, M.; Zgarbová, M.; Jurečka, P.; Svozil, D.; Lankaš, F.; Šponer, J. Can We Accurately Describe the Structure of Adenine Tracts in B-DNA? Reference Quantum-Chemical Computations Reveal Overstabilization of Stacking by Molecular Mechanics. *J. Chem. Theory Comput.* **2012**, *8* (7), 2448–2460.

(123) Kabeláč, M.; Valdes, H.; Sherer, E. C.; Cramer, C. J.; Hobza, P. Benchmark RI-MP2 Database of Nucleic Acid Base Trimers: Performance of Different Density Functional Models for Prediction of Structures and Binding Energies. *Phys. Chem. Chem. Phys.* **2007**, *9* (36), 5000–5008.

(124) Temelso, B.; Archer, K. A.; Shields, G. C. Benchmark Structures and Binding Energies of Small Water Clusters with Anharmonicity Corrections. *J. Phys. Chem. A* **2011**, *115* (43), 12034–12046.

(125) Mas, E. M.; Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; Wormer, P. E. S.; Van Der Avoird, A. Water Pair Potential of near Spectroscopic Accuracy. I. Analysis of Potential Surface and Virial Coefficients. *J. Chem. Phys.* **2000**, *113* (16), 6687–6701.

(126) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; Van Der Avoird, A. Predictions of the Properties of Water from First Principles. *Science* **2007**, *315* (5816), 1249–1252.

(127) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; Van Der Avoird, A. Polarizable Interaction Potential for Water from Coupled Cluster Calculations. I. Analysis of Dimer Potential Energy Surface. *J. Chem. Phys.* **2008**, *128* (9), 094313.

(128) Smith, D. G. A.; Patkowski, K. Toward an Accurate Description of Methane Physisorption on Carbon Nanotubes. *J. Phys. Chem. C* **2014**, *118* (1), 544–550.

(129) Smith, D. G. A.; Patkowski, K. Interactions between Methane and Polycyclic Aromatic Hydrocarbons: A High Accuracy Benchmark Study. *J. Chem. Theory Comput.* **2013**, *9* (1), 370–389.

(130) Vogiatzis, K. D.; Klopper, W.; Friedrich, J. Non-Covalent Interactions of CO2 with Functional Groups of Metal-Organic Frameworks from a CCSD(T) Scheme Applicable to Large Systems. *J. Chem. Theory Comput.* **2015**, *11* (4), 1574–1584.

(131) Smith, D. G. A.; Patkowski, K. Benchmarking the CO2 Adsorption Energy on Carbon Nanotubes. *J. Phys. Chem. C* **2015**, *119* (9), 4934–4948.

(132) Li, S.; Smith, D. G. A.; Patkowski, K. An Accurate Benchmark Description of the Interactions between Carbon Dioxide and Polyheterocyclic Aromatic Compounds Containing Nitrogen. *Phys. Chem. Chem. Phys.* **2015**, *17* (25), 16560–16574.

(133) Li, W.; Grimme, S.; Krieg, H.; Möllmann, J.; Zhang, J. Accurate Computation of Gas Uptake in Microporous Organic Molecular Crystals. *J. Phys. Chem. C* **2012**, *116* (16), 8865–8871.

(134) Lao, K. U.; Schäffer, R.; Jansen, G.; Herbert, J. M. Accurate Description of Intermolecular Interactions Involving Ions Using Symmetry-Adapted Perturbation Theory. *J. Chem. Theory Comput.* **2015**, *11* (6), 2473–2486.

(135) Sharapa, D. I.; Genaev, A.; Cavallo, L.; Minenkov, Y. A Robust and Cost-Efficient Scheme for Accurate Conformational Energies of Organic Molecules. *ChemPhysChem* **2018**, *20* (1), 92–102.

(136) Sellers, B. D.; James, N. C.; Gobbi, A. A Comparison of Quantum and Molecular Mechanical Methods to Estimate Strain Energy in Druglike Fragments. *J. Chem. Inf. Model.* **2017**, *57* (6), 1265–1275.

(137) Fogueri, U. R.; Kozuch, S.; Karton, A.; Martin, J. M. L. The Melatonin Conformer Space: Benchmark and Assessment of Wave Function and DFT Methods for a Paradigmatic Biological and Pharmacological Molecule. *J. Phys. Chem. A* **2013**, *117* (10), 2269–2277.

(138) Tahchieva, D. N.; Bakowies, D.; Ramakrishnan, R.; Von Lilienfeld, O. A. Torsional Potentials of Glyoxal, Oxalyl Halides, and Their Thiocarbonyl Derivatives: Challenges for Popular Density Functional Approximations. *J. Chem. Theory Comput.* **2018**, *14* (9), 4806–4817.

(139) Folmsbee, D.; Hutchison, G. Assessing Conformer Energies Using Electronic Structure and Machine Learning Methods. *Int. J. Quantum Chem.* **2021**, *121* (1), e26381.

(140) Prasad, V. K.; Otero-de-la-Roza, A.; DiLabio, G. A. PEPCONF, A Diverse Data Set of Peptide Conformational Energies. *Sci. Data* **2019**, *6*, 180310.

(141) Goerigk, L.; Karton, A.; Martin, J. M. L.; Radom, L. Accurate Quantum Chemical Energies for Tetrapeptide Conformations: Why MP2 Data with an Insufficient Basis Set Should Be Handled with Caution. *Phys. Chem. Chem. Phys.* **2013**, *15* (19), 7028.

(142) Valdes, H.; Pluháčková, K.; Pitonák, M.; Řezáč, J.; Hobza, P. Benchmark Database on Isolated Small Peptides Containing an Aromatic Side Chain: Comparison between Wave Function and Density Functional Theory Methods and Empirical Force Field. *Phys. Chem. Chem. Phys.* **2008**, *10* (19), 2747.

(143) Kesharwani, M. K.; Karton, A.; Martin, J. M. L. Benchmark *Ab Initio* Conformational Energies for the Proteinogenic Amino Acids through Explicitly Correlated Methods. Assessment of Density Functional Methods. *J. Chem. Theory Comput.* **2016**, *12* (1), 444–454.

(144) Mládek, A.; Krepl, M.; Svozil, D.; Čech, P.; Otyepka, M.; Banáš, P.; Zgarbová, M.; Jurečka, P.; Šponer, J. Benchmark Quantum-Chemical Calculations on a Complete Set of Rotameric Families of the DNA Sugar-Phosphate Backbone and Their Comparison with Modern Density Functional Theory. *Phys. Chem. Chem. Phys.* **2013**, *15* (19), 7295–7310.

(145) Mládek, A.; Banáš, P.; Jurečka, P.; Otyepka, M.; Zgarbová, M.; Šponer, J. Energies and 2′-Hydroxyl Group Orientations of RNA Backbone Conformations. Benchmark CCSD(T)/CBS Database, Electronic Analysis, and Assessment of DFT Methods and MD Simulations. *J. Chem. Theory Comput.* **2014**, *10* (1), 463–480.

(146) Kruse, H.; Mladek, A.; Gkionis, K.; Hansen, A.; Grimme, S.; Sponer, J. Quantum Chemical Benchmark Study on 46 RNA Backbone Families Using a Dinucleotide Unit. *J. Chem. Theory Comput.* **2015**, *11* (10), 4972–4991.

(147) Csonka, G. I.; French, A. D.; Johnson, G. P.; Stortz, C. A. Evaluation of Density Functionals and Basis Sets for Carbohydrates. *J. Chem. Theory Comput.* **2009**, *5* (4), 679–692.

(148) Chan, B. Aqueous-Phase Conformations of Lactose, Maltose, and Sucrose and the Assessment of Low-Cost DFT Methods with the DSCONF Set of Conformers for the Three Disaccharides. *J. Phys. Chem. A* **2020**, *124* (3), 582–590.

(149) Sameera, W. M. C.; Pantazis, D. A. A Hierarchy of Methods for the Energetically Accurate Modeling of Isomerism in Monosaccharides. *J. Chem. Theory Comput.* **2012**, *8* (8), 2630–2645.

(150) Marianski, M.; Supady, A.; Ingram, T.; Schneider, M.; Baldauf, C. Assessing the Accuracy of Across-the-Scale Methods for Predicting Carbohydrate Conformational Energies for the Examples of Glucose and α-Maltose. *J. Chem. Theory Comput.* **2016**, *12* (12), 6157–6168.

(151) Gruzman, D.; Karton, A.; Martin, J. M. L. Performance of Ab Initio and Density Functional Methods for Conformational Equilibria of C_nH_{2n+2} Alkane Isomers (*n*=4−8). *J. Phys. Chem. A* **2009**, *113* (43), 11974–11983.

(152) Kozuch, S.; Bachrach, S. M.; Martin, J. M. L. Conformational Equilibria in Butane-1,4-Diol: A Benchmark of a Prototypical System with Strong Intramolecular H-Bonds. *J. Phys. Chem. A* **2014**, *118* (1), 293–303.

(153) Martin, J. M. L. What Can We Learn about Dispersion from the Conformer Surface of n-Pentane? *J. Phys. Chem. A*

2013, *117* (14), 3118–3132.

(154) Temelso, B.; Klein, K. L.; Mabey, J. W.; Pérez, C.; Pate, B. H.; Kisiel, Z.; Shields, G. C. Exploring the Rich Potential Energy Surface of (H2O)11 and Its Physical Implications. *J. Chem. Theory Comput.* **2018**, *14* (2), 1141–1153.

(155) Morgante, P.; Peverati, R. ACCDB: A Collection of Chemistry Databases for Broad Computational Purposes. *J. Comput. Chem.* **2019**, *40* (6), 839–848.

(156) Yu, H. S.; He, X.; Li, S. L.; Truhlar, D. G. MN15: A Kohn–Sham Global-Hybrid Exchange–Correlation Density Functional with Broad Accuracy for Multi-Reference and Single-Reference Systems and Noncovalent Interactions. *Chem. Sci.* **2016**, *7* (8), 5032–5051.

(157) Yu, L. J.; Karton, A. Assessment of Theoretical Procedures for a Diverse Set of Isomerization Reactions Involving Double-Bond Migration in Conjugated Dienes. *Chem. Phys.* **2014**, *441*, 166–177.

(158) Friedrich, J.; Hänchen, J. Incremental CCSD(T)(F12*)|MP2: A Black Box Method To Obtain Highly Accurate Reaction Energies. *J. Chem. Theory Comput.* **2013**, *9* (12), 5381–5394.

(159) Friedrich, J. Efficient Calculation of Accurate Reaction Energies—Assessment of Different Models in Electronic Structure Theory. *J. Chem. Theory Comput.* **2015**, *11* (8), 3596–3609.

(160) Krieg, H.; Grimme, S. Thermochemical Benchmarking of Hydrocarbon Bond Separation Reaction Energies: Jacob's Ladder Is Not Reversed! *Mol. Phys.* **2010**, *108* (19–20), 2655–2666.

(161) Goerigk, L.; Grimme, S. Efficient and Accurate Double-Hybrid-Meta-GGA Density Functionals—Evaluation with the Extended GMTKN30 Database for General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions. *J. Chem. Theory Comput.* **2011**, *7* (2), 291–309.

(162) Zhao, Y.; González-Garda, N.; Truhlar, D. G. Benchmark Database of Barrier Heights for Heavy Atom Transfer, Nucleophilic Substitution, Association, and Unimolecular Reactions and Its Use to Test Theoretical Methods. *J. Phys. Chem. A* **2005**, *109* (9), 2012–2018.

(163) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. Multi-Coefficient Extrapolated Density Functional Theory for Thermochemistry and Thermochemical Kinetics. *Phys. Chem. Chem. Phys.* **2005**, *7* (1), 43–52.

(164) Goerigk, L.; Grimme, S. A General Database for Main Group Thermochemistry, Kinetics, and Noncovalent Interactions − Assessment of Common and Reparameterized (Meta-)GGA Density Functionals. *J. Chem. Theory Comput.* **2010**, *6* (1), 107–126.

(165) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and Density Functional Theories for the Computation of Enthalpies of Formation. *J. Chem. Phys.* **1997**, *106* (3), 1063–1079.

(166) Yu, L.-J.; Sarrami, F.; O'Reilly, R. J.; Karton, A. Can DFT and Ab Initio Methods Describe All Aspects of the Potential Energy Surface of Cycloreversion Reactions? *Mol. Phys.* **2015**, *114* (1), 21–33.

(167) Karton, A.; Schreiner, P. R.; Martin, J. M. L. Heats of Formation of Platonic Hydrocarbon Cages by Means of High-Level Thermochemical Procedures. *J. Comput. Chem.* **2016**, *37* (1), 49–58.

(168) Karton, A.; Gruzman, D.; Martin, J. M. L. Benchmark Thermochemistry of the CnH2n+2 Alkane Isomers (n = 2−8) and Performance of DFT and Composite Ab Initio Methods for Dispersion-Driven Isomeric Equilibria. *J. Phys. Chem. A* **2009**, *113* (29), 8434–8447.

(169) Johnson, E. R.; Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Delocalization Errors in Density Functionals and Implications for Main-Group Thermochemistry. *J. Chem. Phys.* **2008**, *129* (20), 204112.

(170) Karton, A.; Martin, J. M. L. Explicitly Correlated Benchmark Calculations on C8H8 Isomer Energy Separations: How Accurate Are DFT, Double-Hybrid, and Composite Ab Initio Procedures? *Mol. Phys.* **2012**, *110* (19–20), 2477–2491.

(171) Manna, D.; Martin, J. M. L. What Are the Ground State Structures of C20 and C24? An Explicitly Correlated Ab Initio Approach. *J. Phys. Chem. A* **2015**, *120* (1), 153–160.

(172) Zhao, Y.; Tishchenko, O.; Gour, J. R.; Li, W.; Lutz, J. J.; Piecuch, P.; Truhlar, D. G. Thermochemical Kinetics for Multireference Systems: Addition Reactions of Ozone. *J. Phys. Chem. A* **2009**, *113* (19), 5786–5799.

(173) Lee, J. S. Accurate Ab Initio Binding Energies of Alkaline Earth Metal Clusters. *J. Phys. Chem. A* **2005**, *109* (51), 11927–11932.

(174) Lepetit, C.; Chermette, H.; Gicquel, M.; Heully, J-L.; Chauvin, R. Description of Carbo-Oxocarbons and Assessment of Exchange-Correlation Functionals for the DFT Description of Carbo-Mers. *J. Phys. Chem. A* **2006**, *111* (1), 136–149.

(175) Schreiner, P. R.; Fokin, A. A.; Pascal Jr., R. A.; de Meijere, A. Many Density Functional Theory Approaches Fail To Give Reliable Large Hydrocarbon Isomer Energy Differences. *Org. Lett.* **2006**, *8* (17), 3635–3638.

(176) Woodcock, H. L.; Schaefer III, H. F.; Schreiner, P. R. Problematic Energy Differences between Cumulenes and Poly-Ynes: Does This Point to a Systematic Improvement of Density Functional Theory? *J. Phys. Chem. A* **2002**, *106* (49),

11923–11931.

(177) Piacenza, M.; Grimme, S. Systematic Quantum Chemical Study of DNA-Base Tautomers. *J. Comput. Chem.* **2004**, *25* (1), 83–99.

(178) Grimme, S.; Mück-Lichtenfeld, C.; Würthwein, E.-U.; Ehlers, A. W.; Goumans, T. P. M.; Lammertsma, K. Consistent Theoretical Description of 1,3-Dipolar Cycloaddition Reactions. *J. Phys. Chem. A* **2006**, *110* (8), 2583–2586.

(179) Grimme, S. Semiempirical Hybrid Density Functional with Perturbative Second-Order Correlation. *J. Chem. Phys.* **2006**, *124* (3), 034108.

(180) Karton, A.; O'Reilly, R. J.; Radom, L. Assessment of Theoretical Procedures for Calculating Barrier Heights for a Diverse Set of Water-Catalyzed Proton-Transfer Reactions. *J. Phys. Chem. A* **2012**, *116* (16), 4211–4221.

(181) Grimme, S.; Kruse, H.; Goerigk, L.; Erker, G. The Mechanism of Dihydrogen Activation by Frustrated Lewis Pairs Revisited. *Angew. Chemie Int. Ed.* **2010**, *49* (8), 1402–1405.

(182) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions Based on Quantum Chemistry. *Sci. Data* **2020**, *7* (1), 1–8.

(183) Rudorff, G. F. von; Heinen, S. N.; Bragato, M.; Lilienfeld, O. A. von. Thousands of Reactants and Transition States for Competing E2 and S2 Reactions. *Mach. Learn. Sci. Technol.* **2020**, *1* (4), 045026.

(184) Karton, A.; Goerigk, L. Accurate Reaction Barrier Heights of Pericyclic Reactions: Surprisingly Large Deviations for the CBS-QB3 Composite Method and Their Consequences in DFT Benchmark Studies. *J. Comput. Chem.* **2015**, *36* (9), 622–632.

(185) Zheng, J.; Zhao, Y.; Truhlar, D. G. Representative Benchmark Suites for Barrier Heights of Diverse Reaction Types and Assessment of Electronic Structure Methods for Thermochemical Kinetics. **2007**, *3* (2), 569–582.

(186) Karton, A.; Tarnopolsky, A.; Lamère, J.-F.; Schatz, G. C.; Martin, J. M. L. Highly Accurate First-Principles Benchmark Data Sets for the Parametrization and Validation of Density Functional and Other Approximate Methods. Derivation of a Robust, Generally Applicable, Double-Hybrid Functional for Thermochemistry and Thermochemical Kinetics. *J. Phys. Chem. A* **2008**, *112* (50), 12868–12886.

(187) Goerigk, L.; Sharma, R. The INV24 Test Set: How Well Do Quantum-Chemical Methods Describe Inversion and Racemization Barriers? *Can. J. Chem.* **2016**, *94* (12), 1133–1143.

(188) Yu, L. J.; Sarrami, F.; O'Reilly, R. J.; Karton, A. Reaction Barrier Heights for Cycloreversion of Heterocyclic Rings: An Achilles' Heel for DFT and Standard Ab Initio Procedures. *Chem. Phys.* **2015**, *458*, 1–8.

(189) Karton, A.; O'Reilly, R. J.; Chan, B.; Radom, L. Determination of Barrier Heights for Proton Exchange in Small Water, Ammonia, and Hydrogen Fluoride Clusters with G4(MP2)-Type, MPn, and SCS-MPn Procedures-a Caveat. *J. Chem. Theory Comput.* **2012**, *8* (9), 3128–3136.

(190) Chan, B.; Gilbert, A. T. B.; Gill, P. M. W.; Radom, L. Performance of Density Functional Theory Procedures for the Calculation of Proton-Exchange Barriers: Unusual Behavior of M06-Type Functionals. *J. Chem. Theory Comput.* **2014**, *10* (9), 3777–3783.

(191) Smith, B. J.; Swanton, D. J.; Pople, J. A.; Schaefer III, H. F.; Radom, L. Transition Structures for the Interchange of Hydrogen Atoms within the Water Dimer. *J. Chem. Phys.* **1990**, *92* (2), 1240–1247.

(192) Tschumper, G. S.; Leininger, M. L.; Hoffman, B. C.; Valeev, E. F.; Schaefer III, H. F.; Quack, M. Anchoring the Water Dimer Potential Energy Surface with Explicitly Correlated Computations and Focal Point Analyses. *J. Chem. Phys.* **2002**, *116* (2), 690–701.

(193) Grimme, S.; Steinmetz, M.; Korth, M. How to Compute Isomerization Energies of Organic Molecules with Quantum Chemical Methods. *J. Org. Chem.* **2007**, *72* (6), 2118–2126.

(194) Huenerbein, R.; Schirmer, B.; Moellmann, J.; Grimme, S. Effects of London Dispersion on the Isomerization Reactions of Large Organic Molecules: A Density Functional Benchmark Study. *Phys. Chem. Chem. Phys.* **2010**, *12* (26), 6940–6948.

(195) Grimme, S. Seemingly Simple Stereoelectronic Effects in Alkane Isomers and the Implications for Kohn–Sham Density Functional Theory. *Angew. Chemie Int. Ed.* **2006**, *45* (27), 4460–4464.

(196) Schwabe, T.; Grimme, S. Double-Hybrid Density Functionals with Long-Range Dispersion Corrections: Higher Accuracy and Extended Applicability. *Phys. Chem. Chem. Phys.* **2007**, *9* (26), 3397–3406.

(197) Yu, L.-J.; Sarrami, F.; Karton, A.; O'Reilly, R. J. An Assessment of Theoretical Procedures for π-Conjugation Stabilisation Energies in Enones. *Mol. Phys.* **2014**, *113* (11), 1284–1296.

(198) Karton, A. How Reliable Is DFT in Predicting Relative Energies of Polycyclic Aromatic Hydrocarbon Isomers? Comparison of Functionals from Different Rungs of Jacob's Ladder. *J. Comput. Chem.* **2017**, *38* (6), 370–382.

(199) Yoo, S.; Aprà, E.; Zeng, X. C.; Xantheas, S. S. High-Level Ab Initio Electronic Structure Calculations of Water Clusters (H2O)16 and (H2O)17: A New Global Minimum for (H2O)16. *J. Phys. Chem. Lett.* **2010**, *1* (20), 3122–3127.

(200) Lao, K. U.; Herbert, J. M. Accurate and Efficient Quantum Chemistry Calculations for Noncovalent Interactions in Many-Body Systems: The XSAPT Family of Methods. *J. Phys. Chem. A.* **2015**, *119* (2), 235–252.

(201) Mardirossian, N.; Lambrecht, D. S.; McCaslin, L.; Xantheas, S. S.; Head-Gordon, M. The Performance of Density Functionals for Sulfate-Water Clusters. *J. Chem. Theory Comput.* **2013**, *9* (3), 1368–1380.

(202) Karton, A.; Sylvetsky, N.; Martin, J. M. L. W4-17: A Diverse and High-Confidence Dataset of Atomization Energies for Benchmarking High-Level Electronic Structure Methods. *J. Comput. Chem.* **2017**, *38* (24), 2063–2075.

(203) Taylor, D. E.; Ángyán, J. G.; Galli, G.; Zhang, C.; Gygi, F.; Hirao, K.; Song, J. W.; Rahul, K.; Anatole Von Lilienfeld, O.; Podeszwa, R.; et al. Blind Test of Density-Functional-Based Methods on Intermolecular Interaction Energies. *J. Chem. Phys.* **2016**, *145* (12), 124105.

(204) Donchev, A. G.; Taube, A. G.; Decolvenaere, E.; Hargus, C.; McGibbon, R. T.; Law, K.-H.; Gregersen, B. A.; Li, J.-L.; Palmo, K.; Siva, K.; et al. Quantum Chemical Benchmark Databases of Gold-Standard Dimer Interaction Energies. *Sci. Data* **2021**, *8* (1), 1–9.

(205) Sparrow, Z. M.; Ernst, B. G.; Joo, P. T.; Lao, K. U.; DiStasio Jr., R. A. NENCI-2021 Part I: A Large Benchmark Database of Non-Equilibrium Non-Covalent Interactions Emphasizing Close Intermolecular Contacts. *arXiv* **2021**. (arXiv:2102.02354v1)

(206) Romero-Montalvo, E.; DiLabio, G. A. Computational Study of Hydrogen Bond Interactions in Water Cluster-Organic Molecule Complexes. *J. Phys. Chem. A.* **2021**, *125* (16), 3369–3377.

(207) Miriyala, V. M.; Řezáč, J. Testing Semiempirical Quantum Mechanical Methods on a Data Set of Interaction Energies Mapping Repulsive Contacts in Organic Molecules. *J. Phys. Chem. A* **2018**, *122* (10), 2801–2808.

(208) Kříž, K.; Nováček, M.; Řezáč, J. Non-Covalent Interactions Atlas Benchmark Data Sets 3: Repulsive Contacts. *J. Chem. Theory Comput.* **2021**, *17* (3), 1548–1561.

(209) Mehta, N.; Fellowes, T.; White, J.; Goerigk, L. The CHAL336 Benchmark Set: How Well Do Quantum-Chemical Methods Describe Chalcogen-Bonding Interactions? *J. Chem. Theory Comput.* **2021**, *17* (5), 2783–2806.

(210) Oliveira, V.; Kraka, E.; Cremer, D. The Intrinsic Strength of the Halogen Bond: Electrostatic and Covalent Contributions Described by Coupled Cluster Theory. *Phys. Chem. Chem. Phys.* **2016**, *18* (48), 33031–33046.

(211) Sedlak, R.; Janowski, T.; Pitoňák, M.; Řezáč, J.; Pulay, P.; Hobza, P. Accuracy of Quantum Chemical Methods for Large Noncovalent Complexes. *J. Chem. Theory Comput.* **2013**, *9* (8), 3364–3374.

(212) Calbo, J.; Ortí, E.; Sancho-García, J. C.; Aragó, J. Accurate Treatment of Large Supramolecular Complexes by Double-Hybrid Density Functionals Coupled with Nonlocal van Der Waals Corrections. *J. Chem. Theory Comput.* **2015**, *11* (3), 932–939.

(213) Sure, R.; Grimme, S. Comprehensive Benchmark of Association (Free) Energies of Realistic Host–Guest Complexes. *J. Chem. Theory Comput.* **2015**, *11* (8), 3785–3801.

(214) Ni, Z.; Guo, Y.; Neese, F.; Li, W.; Li, S. Cluster-in-Molecule Local Correlation Method with an Accurate Distant Pair Correction for Large Systems. *J. Chem. Theory Comput.* **2021**, *17* (2), 756–766.

(215) Sharapa, D. I.; Margraf, J. T.; Hesselmann, A.; Clark, T. Accurate Intermolecular Potential for the C60 Dimer: The Performance of Different Levels of Quantum Theory. *J. Chem. Theory Comput.* **2017**, *13* (1), 274–285.

(216) Lao, K. U.; Herbert, J. M. An Improved Treatment of Empirical Dispersion and a Many-Body Energy Decomposition Scheme for the Explicit Polarization plus Symmetry-Adapted Perturbation Theory (XSAPT) Method. *J. Chem. Phys.* **2013**, *139* (3), 034107.

(217) Mezei, P. D.; Csonka, G. I.; Ruzsinszky, A.; Sun, J. Accurate, Precise, and Efficient Theoretical Methods to Calculate Anion-π Interaction Energies in Model Structures. *J. Chem. Theory Comput.* **2015**, *11* (1), 360–371.

(218) Zahn, S.; Macfarlane, D. R.; Izgorodina, E. I. Assessment of Kohn-Sham Density Functional Theory and Møller-Plesset Perturbation Theory for Ionic Liquids. *Phys. Chem. Chem. Phys.* **2013**, *15* (32), 13664–13675.

(219) Zhang, H.; Krupa, J.; Wierzejewska, M.; Biczysko, M. The Role of Dispersion and Anharmonic Corrections in Conformational Analysis of Flexible Molecules: The Allyl Group Rotamerization of Matrix Isolated Safrole. *Phys. Chem. Chem. Phys.* **2019**, *21* (16), 8352–8364.

(220) Kirschner, K. N.; Heiden, W.; Reith, D. Small Alcohols Revisited: CCSD(T) Relative Potential Energies for the Minima, First- and Second-Order Saddle Points, and Torsion-Coupled Surfaces. *ACS Omega* **2018**, *3* (1), 419–432.

(221) Greenwell, C.; Beran, G. J. O. Inaccurate Conformational Energies Still Hinder Crystal Structure Prediction in Flexible Organic Molecules. *Cryst. Growth Des.* **2020**, *20* (8), 4875–4881.

(222) Lahey, S. L. J.; Thien Phuc, T. N.; Rowley, C. N. Benchmarking Force Field and the ANI Neural Network Potentials for the Torsional Potential Energy Surface of Biaryl Drug Fragments. *J. Chem. Inf. Model.* **2020**, *60* (12), 6258–6268.

(223) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x Data Sets, Coupled-Cluster and Density Functional Theory Properties for Molecules. *Sci. Data* **2020**, *7* (1), 1–10.

(224) Řezáč, J.; Bím, D.; Gutten, O.; Rulíšek, L. Toward Accurate Conformational Energies of Smaller Peptides and Medium-Sized Macrocycles: MPCONF196 Benchmark Energy Data Set. *J. Chem. Theory Comput.* **2018**, *14* (3), 1254–1266.

Part VI

Chapter 10

Conclusion

This dissertation presented the development of atom-centered potentials (ACPs) for use with low-cost quantum mechanical (QM) methods to minimize their errors and allow the modeling of large molecular systems. ACPs were specifically designed to yield corrections based on the wavefunction of the underlying method such as small basis set Hartree–Fock (HF) or density-functional theory (DFT) methods. The intent was to reproduce the accuracy in predicting specific molecular properties as obtained with nearly complete basis set wavefunction theory methods but at a computational cost that is orders of magnitude less expensive. The developed ACP based approaches are suitable for modeling systems with many hundreds to a few thousand atoms and ideal for applications such as fast geometry optimizations (local minima and transition states), high-throughput conformer screening, prediction of non-covalent interaction strengths, and modeling of chemical reactions.

The research work presented in this dissertation began with a proof-of-concept study in which the ability of ACPs to correct the underlying deficiencies in a minimal basis set HF method was undertaken. ACPs were developed for four elements (H, C, N, O) and the training set contained a collection of non-covalent properties (9,814 entries), most of which were obtained using highly accurate wavefunction theory methods at the complete basis set limit. The ACPs were optimized using a regularized least-squares regression technique, which is suitable for developing ACPs on much large training sets (Chapters 7–9). The results showed that using ACPs with minimal basis set HF reduced the errors in non-covalent properties by (on average) 67% and with 1.4 kcal/mol within the reference values, with little-to no increase in computational effort relative to the underlying method. These findings validated and supported the hypothesis that ACPs are an effective, inexpensive approach to correcting the deficiencies in quantum mechanical methods.

In order to extend the ACPs to a broader range of chemical properties, it was necessary to develop reference data that were unavailable in the literature. Chapters 4–6 detailed the generation of new, diverse data sets of four molecular properties, including polypeptide conformational energies (PEPCONF), bond separation energies (BSE49), barrier heights (BH9), and reaction energies (BH9-RE). These data sets were used for the ACP development efforts described in Chapters 7–9. The new data sets generated as part of this dissertation will support the development and testing of future generations of ACPs and, more broadly, other computational chemistry methods.

Next, the study in Chapter 7 was aimed at better understanding the extent to which the methodological shortcomings of small basis set HF methods can be mitigated by using ACPs or semi-empirical correction schemes from the literature. Both these corrections are specifically developed to correct the absence of dispersion in HF plus basis set incompleteness effects. In this work, new ACPs were developed with the purpose of comparing their performance to semi-empirical correction schemes from the literature, and were targeted for ten elements (H, B, C, N, O, F, Si, P, S, Cl). They were trained using an extensive set of non-covalent properties (105,880 entries). It was found that small basis set HF used in conjunction with ACPs led to the prediction of non-covalent properties that were generally more accurate than those predicted using previous semi-empirical correction schemes. It was also found that the ACPs could be designed for use with one or more of these correction schemes to reduce the underlying errors even further.

Chapters 8 and 9 presented the development of ACPs designed to mitigate the errors in small basis set HF and DFT methods, respectively. The ACPs were parametrized for ten elements common in organic and bio-organic chemistry (H, C, N, O, F, P, S, Cl) plus boron and silicon using the regularized least-squares regression technique used in the proof-of-concept study. A large and diverse training set was utilized to parametrize these ACPs: 73,832 data points for ACPs with HF methods and 118,655 data points for ACPs with DFT methods. The ACPs were also validated on properties not used in their development – 32,048 data points for HF methods and 42,567 data points for DFT methods.

In Chapter 8, four new sets of ACPs were developed for small basis set HF methods, including those that could be paired with the semi-empirical correction schemes examined in Chapter 7. It should be noted that the target molecular properties for the ACPs developed in Chapter 8 were restricted to those dependent on non-covalent interactions. This is because small basis set HF methods, regardless of the quality of the ACPs or other semi-empirical correction techniques, cannot accurately predict covalent properties like barrier heights of chemical reactions. The assessment of the uncorrected and ACP-corrected small basis set HF methods showed that ACPs can reduce errors in various non-covalent property subsets of the training set by 16–90%. The good performance of ACPs for the training set is also carried over to the validation set with approximately the same performance in terms of average error (error reductions up to 98%). More importantly, the average errors are similar in the training and validation sets, confirming the robustness and applicability of these methods outside the boundaries of the training set. All these findings suggest that the ACPs developed in Chapter 8 can be applied successfully in workflows requiring fast geometry optimizations of large chemical structures, high-throughput conformer screening, and prediction of non-covalent interaction strengths in large systems.

Chapter 9 extended the applicability of ACPs to thermochemical properties where bond breaking and formation occur, such as reaction energies, barrier heights, and bond separation energies. For this purpose, three new sets of ACPs were developed for double-ζ basis set DFT methods. The assessment of the uncorrected and ACP-corrected double-ζ DFT methods showed that ACPs reduce errors in the training set by 16–90% and by 2–50% in the validation set. The similarity in the performance for systems not included in the training set, similar to what was observed for the ACPs in Chapter 8, confirmed the applicability of the ACPs developed for DFT based methods also outside the boundaries of the training set. The findings in Chapter 9 indicate that the developed ACPs for double-ζ basis set DFT methods are more useful to model properties beyond non-covalent interactions and can be successfully applied to tasks involving reaction profile analysis of large systems and fast transition state searches.

Overall, the research studies in this dissertation underscored some of the important successes associated with an ACP based approach. However, certain limitations of an ACP based approach were also identified:

i. *ACP development requires large, diverse, and accurate training sets of reference data*. It was demonstrated in Chapters 7–9 that the properties of systems that were not well-represented in the training set, like chalcogen bonding and repulsive interactions, either improved to a lesser extent or got worse than well-represented properties in the training set. This observation suggests including more diverse systems in the training set for greater robustness of the resulting ACP based methods. However, there is a scarcity of highly accurate reference data in the literature, and the computational costs associated with generating such data are very high.

ii. *The generalization of an ACP based method necessitates the development of ACPs for each element in the periodic table.* The fact that each element should also be represented diversely in the training set is a challenge. It requires collecting reference data of molecular properties involving target periodic table elements, which is generally scarce in the literature or computationally very expensive to generate accurately. Therefore, future work will focus on generating new reference data and on its use in developing ACPs for more elements (particularly alkali, alkaline, and transition metals for modeling metalloproteins and organometallic chemistry). Meanwhile, in the absence of ACPs for all elements, it has already been demonstrated that even if ACPs are only applied to a subset of the atoms in the system, their effect seems to reduce the errors relative to uncorrected methods.

iii. *Larger basis sets are required to predict the properties of some challenging systems accurately*. In the various chapters of this dissertation, it was demonstrated that the developed ACPs are capable of mitigating the error due to basis set incompleteness of small basis sets. However, a limitation remains in describing anion-containing systems caused by the lack of diffuse functions in the basis sets. Therefore, new ACPs developed for small basis sets that include some diffuse functions will be an important area of exploration to overcome the issue of modeling anion-containing systems. Instead of developing ACPs for use with the same basis sets with additional diffuse functions for all elements of interest, a computationally efficient alternative would also be to develop ACPs for use with a mixed basis set technique. This would involve using small basis sets with additional diffuse functions to represent only certain elements which are more likely to be present in an anionic form (like N, O, and S), while the other elements usually present in majority (like H and C) are represented with small basis sets with no additional diffuse functions.

iv. *The performance of ACPs is dependent on the underlying approach and target properties.* In Chapters 7–9, it was demonstrated that the extent to which ACPs can reduce the errors due to approximations in HF or DFT functionals and incompleteness of the basis set relies heavily on the chosen method/basis-set combination and target molecular properties. For example, the performance of small basis set HF is known to be significantly worse when predicting barrier heights and reaction energies. For this reason, ACPs were not developed for thermochemical properties in the case of small basis set HF methods in Chapter 8. Nevertheless, the value of ACPs lies in their flexibility to be developed for any method and basis set combination, which has decent performance for a collection of target molecular properties. Although, a performance depreciation in some properties due to ACPs can be observed when multiple properties with varying magnitudes of reference data are included in the training set compared to when ACPs are developed to be property specific. More generally, further investigation needs to be carried out to understand how ACPs impact the calculation of other properties if included in the training set, such as excitation energies, solvation energies, acid dissociation constants (pK_a), NMR chemical shifts, etc.

v. *ACPs can be used with only selected quantum chemistry packages.* The ACPs developed in this dissertation are only applicable with software that implements effective-core potentials and allows Gaussian-type functions. Besides, the computational cost of using ACPs is directly dependent on the implementation and efficiency of the effective-core potential module in software packages. For example, the use of ACPs in the Gaussian16 program causes only an increase in the computational time of the underlying method by approximately 10–30%. For this reason, the development of ACPs was carried out with the Gaussian16 program throughout this dissertation. The use of ACPs with other software packages requires additional testing.

In conclusion, the major outcome of this dissertation has been the development of computationally inexpensive QM methods to model systems with hundreds to thousands of atoms. Newer generations of ACPs were developed for use with HF and DFT methods combined with small basis sets. It was shown that these ACPs could efficiently predict various molecular properties with low average errors relative to nearly complete basis set wavefunction theory methods. More importantly, as the ACPs generate corrections by operating on the wavefunctions produced by the underlying method, the failures, if any, are far from being catastrophic. Therefore, the ACPs developed in Chapters 8 and 9 are anticipated to reliably enable many QM applications such as modeling biological structure-function relationships, allowing faster structure determinations and conformational samplings, and assisting high-throughput *in silico* screening. The ACP based methods also offer a better alternative to uncorrected small basis set HF or DFT methods, often used in computational organic chemistry studies. This dissertation will motivate further applications in the field of supramolecular host-guest complexation, *ab initio* molecular dynamics, protein structure refinement, and others. One specific field of interest for future work is to apply the developed ACPs to investigate mechanisms of various enzyme-catalyzed reactions with biochemical and industrial relevance. Another interesting area for future work will be to explore the usage of ACP based approaches in supporting the development of machine-learning models in chemistry by enabling the rapid generation of the millions of reference data points required for developing these models. This dissertation has demonstrated that ACPs offer a convenient means of mitigating the underlying shortcomings of various low-cost QM methods thereby leading to improvements in their performance without any significant additional computational cost. Therefore, as new developments continually lead to improved and more accurate quantum mechanical approaches (for example, new approximate density functionals that are closer to the exact density functionals and better represent the physical reality) will allow the opportunity to explore avenues of developing new ACPs that can be coupled with such approaches to enhance their performance further. This will further allow to model various applications in chemistry, biochemistry, and materials science where a better understanding is required of how electrons in molecules behave to control their structure, properties, and reactivity.

Appendices

Appendix 1

Supporting Information for Chapter 3

Section S1. Sample input file demonstrating the use of atom-centered potentials in Gaussian software

A general externally specified basis set file named **“minis.gbs”** is defined by adding the keyword **“gen”**, whereas the additional ACPs file **“minis.acp”** is invoked by mentioning the keyword **“pseudo=read”**. For note, the MINIs basis set is not defined in Gaussian 09 as a keyword and hence must be imported. The ACPs are recommended to be used with only HF-D3/MINIs method. The MINIs basis-set file and the ACP file are given below in separate tables.

```
# hf empiricaldispersion=gd3bj  gen  pseudo=read

Sample water dimer input

0  1
O    -0.702196054    -0.056060256     0.009942262
H    -1.022193224     0.846775782    -0.011488714
H     0.257521062     0.042121496     0.005218999
O     2.220871067     0.026716792     0.000620476
H     2.597492682    -0.411663274     0.766744858
H     2.593135384    -0.449496183    -0.744782026

@minis.gbs

@minis.acp
```

Section S2. MINIs basis set file

```
-H     0
S   3   1.00
      7.0340630             0.0704520
      1.0647560             0.4078260
      0.2365590             0.6477520
****
-C     0
S   3   1.00
    153.1722600             0.0707400
     23.0730300             0.3953800
      4.9232900             0.6633110
S   3   1.00
      6.6166120            -0.0813800
      0.5258560             0.5748530
      0.1699580             0.5024130
P   3   1.00
      4.9129200             0.1099310
      0.9976160             0.4627130
      0.2326850             0.6275140
****
-N     0
S   3   1.00
    218.3644900             0.0678700
     32.5988900             0.3902020
      6.9173900              0.6700830
S   3   1.00
      8.9194260             -0.0808900
      0.7061410              0.5672020
      0.2250540              0.5110920
P   3   1.00
      6.5562720              0.1159190
      1.3490790              0.4699580
      0.3122090              0.6184480
****
-O     0
S   3   1.00
    281.8665800              0.0690600
     42.4160000              0.3931590
      9.0956200              0.6656690
S   3   1.00
     11.7893260             -0.0808200
      0.9128940              0.5820900
      0.2866610              0.4971600
P   3   1.00
      8.2741400              0.1242710
      1.7154630              0.4765940
      0.3830130              0.6130440
****
```

Section S3. ACP file for HF-D3/MINIs

```
-O 0
O 2 0
local
9
2   0.01   -0.000168189346768
2   0.02    0.000605653336217
2   0.04   -0.003300079992238
2   0.06    0.011085794495486
2   0.08   -0.008270683802950
2   0.10   -0.005260533189893
2   0.28    0.004913517342985
2   0.80   -0.278518737445427
2   1.00   -0.031961945580595
s
4
2   0.01    0.026098853095437
2   0.02    0.003553491683057
2   0.04    0.039765977315537
2   0.10    0.078267014256277
p
6
2   0.02   -0.023563745925486
2   0.04   -0.000420343336990
2   0.14   -0.006012021553531
2   0.20    0.037601451410523
2   0.26    0.029862884644057
2   0.28    0.020340757275235
-N 0
N 2 0
local
7
2   0.01    0.000059350258122
2   0.02    0.000147888658312
2   0.04   -0.000304660446507
2   0.06    0.000623903745469
2   0.10   -0.008039283170642
```

```
2   0.16    0.016537322662634
2   0.60   -0.143148487065597
s
5
2   0.01    0.002627897827004
2   0.04   -0.039942399660317
2   0.06   -0.070419183662426
2   0.08   -0.030802524687106
2   0.40    0.075656768043155
p
7
2   0.01   -0.013766927725013
2   0.02    0.016129559622077
2   0.04    0.006210179164714
2   0.06    0.062638925569832
2   0.22    0.042967062782451
2   1.00   -0.018327232173165
2   1.20   -0.244316963135191
-C 0
C 2 0
local
9
2   0.01    0.000018616093002
2   0.02    0.000213321422460
2   0.04   -0.002149334676696
2   0.06    0.005167491733013
2   0.10   -0.015555530011563
2   0.16    0.053440047959739
2   0.26   -0.137586692541511
2   0.60    0.231209213429043
2   1.40   -0.803897402592908
s
5
2   0.01   -0.015521341917841
2   0.02    0.000455496502618
2   0.16    0.026929386415710
2   0.24    0.019439254877049
2   0.26    0.031673892758438
p
7
2   0.02    0.011113341550083
2   0.08   -0.012093207739110
2   0.18    0.008520890813297
2   0.20    0.059951110662814
2   0.22    0.013163352582365
2   0.24    0.126648139203848
2   1.20   -0.232183686395611
-H 0
H 1 0
local
9
2   0.01   -0.000019380890887
2   0.02   -0.000033664144254
2   0.04    0.001323631398610
2   0.06   -0.003005639829393
2   0.10    0.000900728930274
2   0.12    0.001415645393693
2   0.22    0.008584107592887
2   0.40   -0.025198082699055
2   1.00    0.030400389243154
s
8
2   0.01    0.008971847271946
2   0.02   -0.039538281441503
```

```
2   0.04    0.048679722150789
2   0.06    0.034499415065515
2   0.10   -0.015906029112128
2   0.14   -0.136749456855188
2   0.40    0.402142301245079
2   2.50   -1.033530058142526
```

Section S4. Formulas for the statistical error measures

a) Mean absolute error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where, $x_i = \left|x_{calc,i} - x_{ref,i}\right|$

b) Mean signed error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where, $x_i = x_{calc,i} - x_{ref,i}$

c) Maximum absolute error (MAXE)

$$MAXE = \max_{i} \left|x_{calc,i} - x_{ref,i}\right|$$

d) Minimum absolute error (MINE)

$$MINE = \min_{i} \left|x_{calc,i} - x_{ref,i}\right|$$

e) Mean absolute percent error (MAPE)

$$MAPE = \frac{100}{n}\sum_{i=1}^{n} \frac{x_i}{\left|x_{ref,i}\right|}$$

where, $x_i = \left|x_{calc,i} - x_{ref,i}\right|$

f) Mean signed percent error (MSPE)

$$MSPE = \frac{100}{n}\sum_{i=1}^{n} \frac{x_i}{x_{ref,i}}$$

where, $x_i = x_{calc,i} - x_{ref,i}$

g) Root-mean-square error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

where, $x_i = x_{calc,i} - x_{ref,i}$

h) Standard deviation (SD)

$$SD = \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

where, $x_i = x_{calc,i} - x_{ref,i}$

$$\overline{x} = \frac{1}{n}(x_{calc,i} - x_{ref,i})$$

i) Uncertainty of the MAE

$$UNC = \frac{\sigma}{\sqrt{n}}$$

Table S1. The statistics with respect to reference data of the various databases present in the training/fitting set. The numbers in bracket in the first column indicate the number of data points. (MAE = mean absolute error in kcal/mol. MSE = mean signed error in kcal/mol. MAXE = maximum absolute error in kcal/mol. MINE = minimum absolute error in kcal/mol. MAPE = maximum absolute percent (%) error. MSPE = maximum signed percent (%) error. RMSE = root-mean- square error in kcal/mol. SD = standard deviation in kcal/mol. UNC = uncertainty of the mean absolute error in kcal/mol.)

Data set (#)		**HF-D3/MINIs**	**HF-D3/aug-cc-pVTZ**	**HF-D3/MINIs-ACP**	**HF-3c**
S22x5 (110)	MAE	1.40	0.70	0.36	0.53
	MSE	-1.18	-0.52	-0.11	0.17
	MAXE	8.26	3.42	1.67	2.62
	MINE	0.00	0.00	0.02	0.00
	MAPE	54.79	28.86	32.94	23.39
	MSPE	46.47	4.41	21.73	7.27
	RMSE	2.42	1.03	0.50	0.80
	SD	2.12	0.89	0.49	0.79
	UNC	0.20	0.09	0.05	0.07
S66x8 (528)	MAE	1.24	0.51	0.28	0.38
	MSE	-1.15	-0.44	-0.16	0.08
	MAXE	6.68	2.69	1.33	2.46
	MINE	0.00	0.00	0.00	0.00
	MAPE	45.66	19.33	29.58	19.61
	MSPE	38.70	7.23	13.84	5.30
	RMSE	1.78	0.72	0.37	0.56
	SD	1.36	0.56	0.33	0.56
	UNC	0.06	0.02	0.01	0.02
ACHC (54)	MAE	1.44	0.34	0.27	0.28
	MSE	-1.44	0.34	-0.13	0.06
	MAXE	5.62	0.87	1.46	1.48
	MINE	0.01	0.01	0.01	0.00

Data set (#)		HF-D3/MINIs	HF-D3/aug-cc-pVTZ	HF-D3/MINIs-ACP	HF-3c
	MAPE	27.91	6.70	6.05	5.68
	MSPE	27.91	-6.62	3.37	-0.69
	RMSE	1.66	0.40	0.37	0.37
	SD	0.83	0.22	0.35	0.36
	UNC	0.11	0.03	0.05	0.05
BBI[a] (94)	MAE	1.05	0.60	0.22	0.87
	MSE	-1.04	-0.50	0.13	0.87
	MAXE	2.33	1.11	1.10	1.79
	MINE	0.02	0.02	0.01	0.15
	MAPE	23.29	10.76	5.37	20.26
	MSPE	22.81	8.17	0.37	-19.05
	RMSE	1.19	0.65	0.32	0.91
	SD	0.58	0.42	0.29	0.25
	UNC	0.06	0.04	0.03	0.03
SSI[a,b] (2449)	MAE	0.84	0.23	0.17	0.21
	MSE	-0.84	-0.20	0.00	-0.11
	MAXE	6.52	1.37	2.49	3.28
	MINE	0.01	0.00	0.00	0.00
	MAPE	154.74	43.57	30.27	29.79
	MSPE	45.86	11.87	-1.09	14.69
	RMSE	0.99	0.29	0.26	0.30
	SD	0.51	0.21	0.26	0.28
	UNC	0.01	0.00	0.01	0.01
P26 (69)	MAE	2.06	0.68	0.47	1.20
	MSE	2.02	-0.19	0.15	0.60
	MAXE	7.36	2.15	1.55	4.30
	MINE	0.08	0.01	0.00	0.03
	MAPE	166.40	59.64	32.96	84.07
	MSPE	160.98	-22.81	13.27	46.65
	RMSE	2.63	0.88	0.59	1.59
	SD	1.70	0.86	0.58	1.49
	UNC	0.20	0.10	0.07	0.18
DIPEPCONF (1530)	MAE	2.44	0.89	0.85	1.15
	MSE	1.82	0.74	0.05	-0.58
	MAXE	10.40	3.44	4.20	5.00
	MINE	0.01	0.00	0.00	0.00
	MAPE	49.09	13.59	16.64	21.40
	MSPE	33.55	10.09	3.84	-3.78
	RMSE	3.06	1.09	1.09	1.44
	SD	2.46	0.80	1.09	1.32
	UNC	0.06	0.02	0.03	0.03
MOLdef (4980)	MAE	3.42	1.73	0.90	2.90
	MSE	2.49	1.34	0.18	2.66
	MAXE	63.62	13.36	14.99	18.74
	MINE	0.00	0.00	0.00	0.00

Data set (#)		HF-D3/MINIs	HF-D3/aug-cc-pVTZ	HF-D3/MINIs-ACP	HF-3c
	MAPE	42.17	22.45	12.68	33.32
	MSPE	23.32	16.61	1.19	27.52
	RMSE	5.23	2.41	1.50	3.89
	SD	4.60	2.00	1.49	2.84
	UNC	0.07	0.03	0.02	0.04

[a)] considering data points consisting of only H, C, N and O atoms exclusively. [b)] only neutral systems present in the set were considered.

Table S2. The statistics with respect to reference data of the various noncovalent interactions databases present in the validation set. The numbers in bracket in the first column indicate the number of data points. (MAE = mean absolute error in kcal/mol. MSE = mean signed error in kcal/mol. MAXE = maximum absolute error in kcal/mol. MINE = minimum absolute error in kcal/mol. MAPE = maximum absolute percentage (%) error. MSPE = maximum signed percentage (%) error. RMSE = root mean squared error in kcal/mol. SD = standard deviation in kcal/mol. UNC = uncertainty of the mean absolute error in kcal/mol.)

Data set (#)		HF-D3/MINIs	HF-D3/MINIs-ACP	HF-3c
HC12 (12)	MAE	1.80	0.26	0.42
	MSE	-1.80	0.01	-0.42
	MAXE	3.39	0.64	0.91
	MINE	0.39	0.02	0.03
	MAPE	58.67	9.54	13.78
	MSPE	58.67	-1.91	13.53
	RMSE	1.96	0.32	0.51
	SD	0.81	0.33	0.30
	UNC	0.23	0.10	0.09
ADIM6 (6)	MAE	1.90	0.21	0.47
	MSE	-1.90	-0.05	-0.47
	MAXE	3.16	0.38	0.71
	MINE	0.67	0.02	0.13
	MAPE	56.54	8.08	13.67
	MSPE	56.54	-2.81	13.67
	RMSE	2.09	0.24	0.52
	SD	0.95	0.26	0.24
	UNC	0.39	0.11	0.10
CH_4·PAH (382)	MAE	1.19	0.29	0.19
	MSE	-1.19	0.03	-0.19
	MAXE	6.94	1.28	1.58
	MINE	0.00	0.00	0.00
	MAPE	184.97	60.51	28.94
	MSPE	107.26	26.43	15.88
	RMSE	1.84	0.35	0.32
	SD	1.40	0.35	0.26
	UNC	0.07	0.02	0.01

Data set (#)		HF-D3/MINIs	HF-D3/MINIs-ACP	HF-3c
$C_2H_4\cdot$NT (75)	MAE	1.71	0.67	0.48
	MSE	-1.70	-0.13	-0.45
	MAXE	5.52	2.82	2.06
	MINE	0.00	0.02	0.00
	MAPE	67.52	27.22	21.96
	MSPE	19.63	-5.90	6.79
	RMSE	2.53	0.95	0.70
	SD	1.89	0.95	0.54
	UNC	0.22	0.11	0.06
A24[a] (19)	MAE	0.73	0.32	0.44
	MSE	-0.73	-0.03	-0.11
	MAXE	2.56	0.74	1.33
	MINE	0.19	0.03	0.00
	MAPE	50.35	26.79	27.75
	MSPE	11.13	-18.85	-13.64
	RMSE	1.01	0.38	0.60
	SD	0.71	0.39	0.60
	UNC	0.16	0.09	0.14
HSG (21)	MAE	1.69	0.62	0.74
	MSE	-1.56	-0.45	-0.18
	MAXE	10.29	3.79	4.23
	MINE	0.32	0.01	0.01
	MAPE	95.22	29.46	19.94
	MSPE	24.87	5.09	5.00
	RMSE	2.96	1.08	1.29
	SD	2.58	1.01	1.31
	UNC	0.56	0.22	0.29
HBC6 (118)	MAE	3.27	0.76	1.13
	MSE	-2.84	-0.36	-0.01
	MAXE	11.34	3.84	4.75
	MINE	0.02	0.01	0.00
	MAPE	28.90	13.94	15.69
	MSPE	12.10	8.17	-7.93
	RMSE	4.65	1.09	1.48
	SD	3.70	1.03	1.48
	UNC	0.34	0.09	0.14
$CO_2\cdot$PAH (249)	MAE	1.64	0.87	0.55
	MSE	-1.59	-0.18	-0.32
	MAXE	9.01	4.09	2.70
	MINE	0.00	0.00	0.00
	MAPE	92.83	112.49	50.93
	MSPE	-10.03	-35.36	-17.89
	RMSE	2.62	1.09	0.82
	SD	2.09	1.08	0.76

Data set (#)		HF-D3/MINIs	HF-D3/MINIs-ACP	HF-3c
	UNC	0.13	0.07	0.05
CO_2·NPHAC (96)	MAE	1.92	0.89	0.63
	MSE	-1.92	-0.66	-0.58
	MAXE	11.42	8.20	6.59
	MINE	0.01	0.01	0.00
	MAPE	78.05	42.49	23.53
	MSPE	15.45	-3.67	-3.58
	RMSE	3.04	1.47	1.19
	SD	2.37	1.32	1.05
	UNC	0.24	0.13	0.11
S12L[a] (10)	MAE	15.39	10.27	6.28
	MSE	-15.39	-4.40	-5.15
	MAXE	25.18	19.37	11.75
	MINE	9.87	3.07	0.61
	MAPE	48.38	30.57	18.37
	MSPE	48.38	9.11	14.69
	RMSE	16.27	11.94	7.62
	SD	5.57	11.70	5.92
	UNC	1.76	3.70	1.87
S30L[a] (23)	MAE	15.54	7.53	6.01
	MSE	-15.54	-4.58	-4.75
	MAXE	41.59	26.92	15.83
	MINE	4.80	0.02	0.21
	MAPE	40.76	19.25	15.49
	MSPE	40.76	9.17	11.27
	RMSE	17.85	10.34	7.39
	SD	8.97	9.48	5.80
	UNC	1.87	1.98	1.21
SHILEDS38 (38)	MAE	30.62	4.99	7.67
	MSE	-30.62	-4.98	-7.67
	MAXE	60.84	10.05	15.61
	MINE	3.54	0.10	0.24
	MAPE	63.99	10.31	15.70
	MSPE	63.99	10.27	15.70
	RMSE	33.29	5.53	8.53
	SD	13.25	2.44	3.77
	UNC	2.15	0.40	0.61
CHARGED (886)	MAE	3.25	1.95	2.41
	MSE	-2.34	-0.21	-0.48
	MAXE	18.25	8.41	12.06
	MINE	0.00	0.00	0.01
	MAPE	27.59	27.91	27.44
	MSPE	9.73	0.72	-1.99
	RMSE	4.75	2.52	3.22

Data set (#)		HF-D3/MINIs	HF-D3/MINIs-ACP	HF-3c
	SD	4.14	2.52	3.19
	UNC	0.14	0.08	0.11

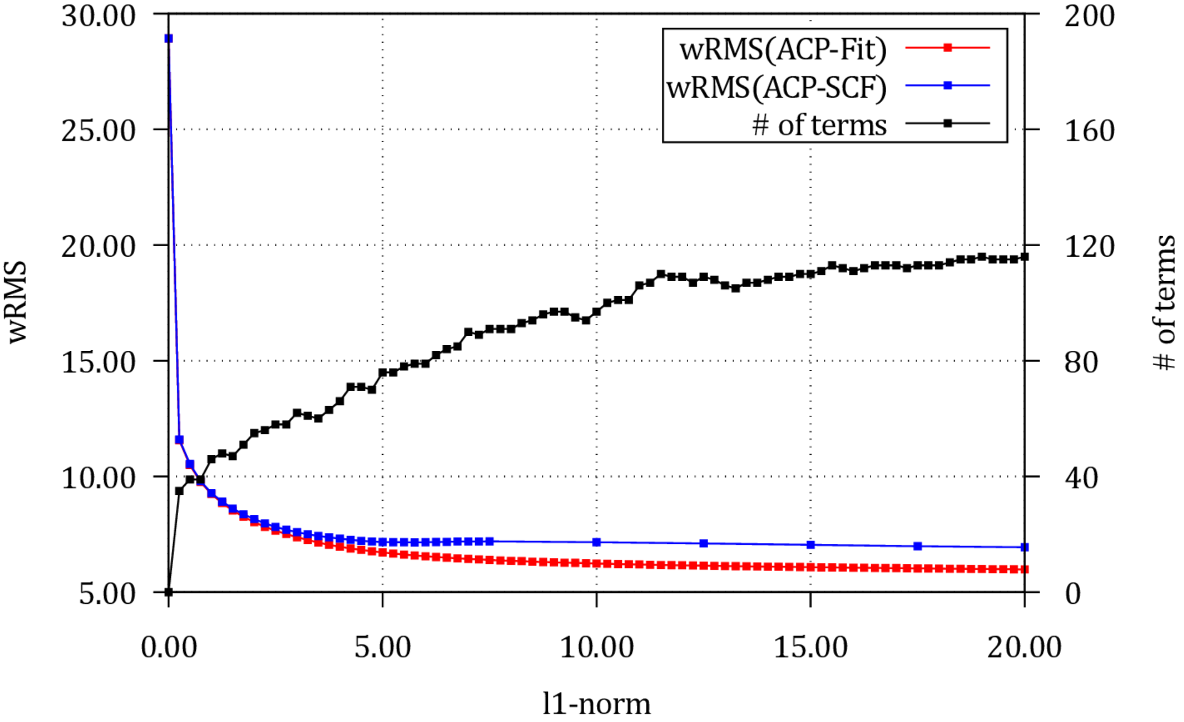
[a] considering data points consisting of only H, C, N and O atoms exclusively.

Table S3. The statistics with respect to reference data of various conformational energy databases present in the validation set. The numbers in bracket in the first column indicate the number of data points. (MAE = mean absolute error in kcal/mol. MSE = mean signed error in kcal/mol. MAXE = maximum absolute error in kcal/mol. MINE = minimum absolute error in kcal/mol. MAPE = maximum absolute percentage (%) error. MSPE = maximum signed percentage (%) error. RMSE = root mean squared error in kcal/mol. SD = standard deviation in kcal/mol. UNC = uncertainty of the mean absolute error in kcal/mol.)

Data set (#)	Statistics	HF-D3/MINIs	HF-D3/MINIs-ACP	HF-3c
ACONF (15)	MAE	1.44	0.98	0.89
	MSE	-1.44	-0.98	-0.89
	MAXE	2.63	1.63	1.76
	MINE	0.67	0.45	0.47
	MAPE	105.34	70.88	69.53
	MSPE	-105.34	-70.88	-69.53
	RMSE	1.55	1.05	0.96
	SD	0.58	0.37	0.36
	UNC	0.15	0.10	0.09
BCONF (64)	MAE	2.40	0.50	0.58
	MSE	2.34	0.36	0.53
	MAXE	3.69	1.52	1.40
	MINE	0.02	0.01	0.01
	MAPE	86.40	22.35	23.48
	MSPE	81.42	9.14	17.34
	RMSE	2.53	0.60	0.69
	SD	0.98	0.49	0.44
	UNC	0.12	0.06	0.06
MCONF (51)	MAE	0.88	0.71	0.89
	MSE	0.44	-0.43	-0.34
	MAXE	2.97	1.47	2.15
	MINE	0.06	0.00	0.00
	MAPE	27.00	16.59	22.37
	MSPE	16.42	-4.75	-2.39
	RMSE	1.19	0.78	1.10
	SD	1.12	0.66	1.06
	UNC	0.16	0.09	0.15
PCONF (10)	MAE	2.43	0.50	2.28
	MSE	2.43	0.49	2.28
	MAXE	4.38	0.88	3.32
	MINE	1.08	0.04	0.11

Data set (#)	Statistics	HF-D3/MINIs	HF-D3/MINIs-ACP	HF-3c
	MAPE	154.09	48.11	195.92
	MSPE	154.09	47.53	195.92
	RMSE	2.65	0.59	2.49
	SD	1.14	0.34	1.06
	UNC	0.36	0.11	0.33
SCONF (17)	MAE	5.20	1.17	1.47
	MSE	1.51	0.02	-0.42
	MAXE	14.76	5.14	7.63
	MINE	0.02	0.02	0.02
	MAPE	152.44	65.47	64.75
	MSPE	-26.72	-41.19	-46.51
	RMSE	6.35	1.65	2.57
	SD	6.35	1.71	2.61
	UNC	1.54	0.41	0.63
TRCONF (8)	MAE	5.14	0.80	3.64
	MSE	-5.14	-0.79	-3.64
	MAXE	11.97	2.88	8.23
	MINE	0.02	0.04	0.09
	MAPE	412.27	49.08	276.78
	MSPE	-412.27	-47.33	-276.78
	RMSE	6.49	1.19	4.45
	SD	4.24	0.95	2.73
	UNC	1.50	0.34	0.96

Figure S1. The combined plot of wRMS vs. l_1-norm and the total number of ACP terms selected by LASSO vs. l_1-norm



The plot illustrates the trend of wRMS (as obtained separately from the LASSO fitting procedure and Gaussian 09-SCF validation) as a function of $l1$-norm of the coefficients. The ACP with $l1$-norm bound of 5.0 a.u. on the coefficients yields an wRMS which is more or less near to the minimum of the wRMS vs. $l1$-norm plot for both the red and blue curves (ACP-fit and ACP-SCF) and the differences of wRMS at that point depicts a minimal non-linearity error. The wRMS of the SCF validation however shows greater deviation from the ACP-fit wRMS as the total number of ACP terms selected by LASSO and correspondingly as the $l1$-norm of the coefficients increases.

Appendix 2

Supporting Information for Chapter 4

The 210 dipeptide sequences considered for this work in terms of their three letter amino acid codes are as follows:

ALA-ALA, ALA-ASN, ALA-CYS, ALA-GLN, ALA-GLY, ALA-HIS, ALA-ILE, ALA-LEU, ALA-MET, ALA-PHE, ALA-PRO, ALA-SER, ALA-THR, ALA-TRP, ALA-TYR, ALA-VAL, ARG-ALA, ARG-ARG, ARG-ASN, ARG-CYS, ARG-GLN, ARG-GLY, ARG-HIS, ARG-ILE, ARG-LEU, ARG-LYS, ARG-MET, ARG-PHE, ARG-PRO, ARG-SER, ARG-THR, ARG-TRP, ARG-TYR, ARG-VAL, ASN-ASN, ASN-CYS, ASN-GLN, ASN-MET, ASP-ALA, ASP-ARG, ASP-ASN, ASP-ASP, ASP-CYS, ASP-GLN, ASP-GLU, ASP-GLY, ASP-HIS, ASP-ILE, ASP-LEU, ASP-LYS, ASP-MET, ASP-PHE, ASP-PRO, ASP-SER, ASP-THR, ASP-TRP, ASP-TYR, ASP-VAL, CYS-CYS, CYS-MET, GLN-CYS, GLN-GLN, GLN-MET, GLU-ALA, GLU-ARG, GLU-ASN, GLU-CYS, GLU-GLN, GLU-GLU, GLU-GLY, GLU-HIS, GLU-ILE, GLU-LEU, GLU-LYS, GLU-MET, GLU-PHE, GLU-PRO, GLU-SER, GLU-THR, GLU-TRP, GLU-TYR, GLU-VAL, GLY-ASN, GLY-CYS, GLY-GLN, GLY-GLY, GLY-HIS, GLY-ILE, GLY-LEU, GLY-MET, GLY-PHE, GLY-PRO, GLY-SER, GLY-THR, GLY-TRP, GLY-TYR, GLY-VAL, HIS-ASN, HIS-CYS, HIS-GLN, HIS-HIS, HIS-MET, HIS-SER, HIS-THR, ILE-ASN, ILE-CYS, ILE-GLN, ILE-HIS, ILE-ILE, ILE-LEU, ILE-MET, ILE-PHE, ILE-PRO, ILE-SER, ILE-THR, ILE-TRP, ILE-TYR, ILE-VAL, LEU-ASN, LEU-CYS, LEU-GLN, LEU-HIS, LEU-LEU, LEU-MET, LEU-PHE, LEU-PRO, LEU-SER, LEU-THR, LEU-TRP, LEU-TYR, LEU-VAL, LYS-ALA, LYS-ASN, LYS-CYS, LYS-GLN, LYS-GLY, LYS-HIS, LYS-ILE, LYS-LEU, LYS-LYS, LYS-MET, LYS-PHE, LYS-PRO, LYS-SER, LYS-THR, LYS-TRP, LYS-TYR, LYS-VAL, MET-MET, PHE-ASN, PHE-CYS, PHE-GLN, PHE-HIS, PHE-MET, PHE-PHE, PHE-SER, PHE-THR, PHE-TRP, PHE-TYR, PRO-ASN, PRO-CYS, PRO-GLN, PRO-HIS, PRO-MET, PRO-PHE, PRO-PRO, PRO-SER, PRO-THR, PRO-TRP, PRO-TYR, PRO-VAL, SER-ASN, SER-CYS, SER-GLN, SER-MET, SER-SER, SER-THR, THR-ASN, THR-CYS, THR-GLN, THR-MET, THR-THR, TRP-ASN, TRP-CYS, TRP-GLN, TRP-HIS, TRP-MET, TRP-SER, TRP-THR, TRP-TRP, TRP-TYR, TYR-ASN, TYR-CYS, TYR-GLN, TYR-HIS, TYR-MET, TYR-SER, TYR-THR, TYR-TYR, VAL-ASN, VAL-CYS, VAL-GLN, VAL-HIS, VAL-MET, VAL-PHE, VAL-SER, VAL-THR, VAL-TRP, VAL-TYR, VAL-VAL.

The 288 tripeptide sequences considered for this work in terms of their three letter amino acid codes are as follows:

GLH-GLH-GLH, GLH-GLN-GLH, GLH-HIS-GLH, GLH-LEU-GLH, GLH-MET-GLH, GLH-PRO-GLH, GLH-TRP-GLH, GLH-TYR-GLH, GLN-GLH-GLH, GLN-GLH-GLN, GLN-GLH-HIS, GLN-GLH-MET, GLN-GLN-GLH, GLN-GLN-GLN, GLN-GLN-HIS, GLN-GLN-MET, GLN-HIS-GLH, GLN-HIS-GLN, GLN-HIS-HIS, GLN-HIS-MET, GLN-LEU-GLH, GLN-LEU-GLN, GLN-LEU-HIS, GLN-LEU-MET, GLN-MET-GLH, GLN-MET-GLN, GLN-MET-HIS, GLN-MET-MET, GLN-PRO-GLH, GLN-PRO-GLN, GLN-PRO-HIS, GLN-PRO-MET, GLN-TRP-GLH, GLN-TRP-GLN, GLN-TRP-HIS, GLN-TRP-MET, GLN-TYR-GLH, GLN-TYR-GLN, GLN-TYR-HIS, GLN-TYR-MET, HIS-GLH-GLH, HIS-GLH-HIS, HIS-GLN-GLH, HIS-GLN-HIS, HIS-HIS-GLH, HIS-HIS-HIS, HIS-LEU-GLH, HIS-LEU-HIS, HIS-MET-GLH, HIS-MET-HIS, HIS-PRO-GLH, HIS-PRO-HIS, HIS-TRP-GLH, HIS-TRP-HIS, HIS-TYR-GLH, HIS-TYR-HIS, LEU-GLH-GLH, LEU-GLH-GLN, LEU-GLH-HIS, LEU-GLH-LEU, LEU-GLH-MET, LEU-GLH-PRO, LEU-GLH-TRP, LEU-GLH-TYR, LEU-GLN-GLH, LEU-GLN-GLN, LEU-GLN-HIS, LEU-GLN-LEU, LEU-GLN-MET, LEU-GLN-PRO, LEU-GLN-TRP, LEU-GLN-TYR, LEU-HIS-GLH, LEU-HIS-GLN, LEU-HIS-HIS, LEU-HIS-LEU, LEU-HIS-MET, LEU-HIS-PRO, LEU-HIS-TRP, LEU-HIS-TYR, LEU-LEU-GLH, LEU-LEU-GLN, LEU-LEU-HIS, LEU-LEU-LEU, LEU-LEU-MET, LEU-LEU-PRO, LEU-LEU-TRP, LEU-LEU-TYR, LEU-MET-GLH, LEU-MET-GLN, LEU-MET-HIS, LEU-MET-LEU, LEU-MET-MET, LEU-MET-PRO, LEU-MET-TRP, LEU-MET-TYR, LEU-PRO-GLH, LEU-PRO-GLN, LEU-PRO-HIS, LEU-PRO-LEU, LEU-PRO-MET, LEU-PRO-PRO, LEU-PRO-TRP, LEU-PRO-TYR, LEU-TRP-GLH, LEU-TRP-GLN, LEU-TRP-HIS, LEU-TRP-LEU, LEU-TRP-MET, LEU-TRP-PRO, LEU-TRP-TRP, LEU-TRP-TYR, LEU-TYR-GLH, LEU-TYR-GLN, LEU-TYR-HIS, LEU-TYR-LEU, LEU-TYR-MET, LEU-TYR-PRO, LEU-TYR-TRP, LEU-TYR-TYR, MET-GLH-GLH, MET-GLH-HIS, MET-GLH-MET, MET-GLN-GLH, MET-GLN-HIS, MET-GLN-MET, MET-HIS-GLH, MET-HIS-HIS, MET-HIS-MET, MET-LEU-GLH, MET-LEU-HIS, MET-LEU-MET, MET-MET-GLH, MET-MET-HIS, MET-MET-MET, MET-PRO-GLH, MET-PRO-HIS, MET-PRO-MET, MET-TRP-GLH, MET-TRP-HIS, MET-TRP-MET, MET-TYR-GLH, MET-TYR-HIS, MET-TYR-MET, PRO-GLH-GLH, PRO-GLH-GLN, PRO-GLH-HIS, PRO-GLH-MET, PRO-GLH-PRO, PRO-GLH-TRP, PRO-GLH-TYR, PRO-GLN-GLH, PRO-GLN-GLN, PRO-GLN-HIS, PRO-GLN-MET, PRO-GLN-PRO, PRO-GLN-TRP, PRO-GLN-TYR, PRO-HIS-GLH, PRO-HIS-GLN, PRO-HIS-HIS, PRO-HIS-MET, PRO-HIS-PRO, PRO-HIS-TRP, PRO-HIS-TYR, PRO-LEU-GLH, PRO-LEU-GLN, PRO-LEU-HIS, PRO-LEU-MET, PRO-LEU-PRO, PRO-LEU-TRP, PRO-LEU-TYR, PRO-MET-GLH, PRO-MET-GLN, PRO-MET-HIS, PRO-MET-MET, PRO-MET-PRO, PRO-MET-TRP, PRO-MET-TYR, PRO-PRO-GLH, PRO-PRO-GLN, PRO-PRO-HIS, PRO-PRO-MET, PRO-PRO-PRO, PRO-PRO-TRP, PRO-PRO-TYR, PRO-TRP-GLH, PRO-TRP-GLN, PRO-TRP-HIS, PRO-TRP-MET, PRO-TRP-PRO, PRO-TRP-TRP, PRO-TRP-TYR, PRO-TYR-GLH, PRO-TYR-GLN, PRO-TYR-HIS, PRO-TYR-MET, PRO-TYR-PRO, PRO-TYR-TRP, PRO-TYR-TYR, TRP-GLH-GLH, TRP-GLH-GLN, TRP-GLH-HIS, TRP-GLH-MET, TRP-GLH-TRP, TRP-GLH-TYR, TRP-GLN-GLH, TRP-GLN-GLN, TRP-GLN-HIS, TRP-GLN-MET, TRP-GLN-TRP, TRP-GLN-TYR, TRP-HIS-GLH, TRP-HIS-GLN, TRP-HIS-HIS, TRP-HIS-MET, TRP-HIS-TRP, TRP-HIS-TYR, TRP-LEU-GLH, TRP-LEU-GLN, TRP-LEU-HIS, TRP-LEU-MET, TRP-LEU-TRP, TRP-LEU-TYR, TRP-MET-GLH, TRP-MET-GLN, TRP-MET-HIS, TRP-MET-MET, TRP-MET-TRP, TRP-MET-TYR, TRP-PRO-GLH, TRP-PRO-GLN, TRP-PRO-HIS, TRP-PRO-MET, TRP-PRO-TRP, TRP-PRO-TYR, TRP-TRP-GLH, TRP-TRP-GLN, TRP-TRP-HIS, TRP-TRP-MET, TRP-TRP-TRP, TRP-TRP-TYR, TRP-TYR-GLH, TRP-TYR-GLN, TRP-TYR-HIS, TRP-TYR-MET, TRP-TYR-TRP, TRP-TYR-TYR, TYR-GLH-GLH, TYR-GLH-GLN, TYR-GLH-HIS, TYR-GLH-MET, TYR-GLH-TYR, TYR-GLN-GLH, TYR-GLN-GLN, TYR-GLN-HIS, TYR-GLN-MET, TYR-GLN-TYR, TYR-HIS-GLH, TYR-HIS-GLN, TYR-HIS-HIS, TYR-HIS-MET, TYR-HIS-TYR, TYR-LEU-GLH, TYR-LEU-GLN, TYR-LEU-HIS, TYR-LEU-MET, TYR-LEU-TYR, TYR-MET-GLH, TYR-MET-GLN, TYR-MET-HIS, TYR-MET-MET, TYR-MET-TYR, TYR-PRO-GLH, TYR-PRO-GLN, TYR-PRO-HIS, TYR-PRO-MET, TYR-PRO-TYR, TYR-TRP-GLH, TYR-TRP-GLN, TYR-TRP-HIS, TYR-TRP-MET, TYR-TRP-TYR, TYR-TYR-GLH, TYR-TYR-GLN, TYR-TYR-HIS, TYR-TYR-MET, TYR-TYR-TYR.

The 154 four-character Protein Data Bank (PDB) codes from which the disulfide-bridged oligopeptides were extracted are as follows:

1a43, 1aum, 1avp, 1baj, 1bmg, 1bvo, 1bwz, 1c8e, 1eha, 1eia, 1emr, 1f02, 1f6l, 1gku, 1ijs, 1jjh, 1k5h, 1kac, 1ml8, 1mqa, 1mry, 1nov, 1ny7, 1nyl, 1ou5, 1p5y, 1pfc, 1plr, 1q7q, 1qb3, 1qe0, 1qfp, 1rlr, 1ry7, 1s0g, 1s94, 1se2, 1t3b, 1tgo, 1tjd, 1tmf, 1vb2, 1ver, 1vkx, 1wcs, 1xyh, 1zmw, 1zzd, 2a1r, 2a1s, 2a8z, 2ayu, 2cas, 2czk, 2duk, 2ecf, 2h2y, 2h4m, 2h4r, 2if9, 2irm, 2lve, 2mha, 2o8v, 2ot8, 2ov8, 2p62, 2q2p, 2q98, 2r30, 2vaj, 2w2s, 2wxw, 2wzr, 2xpe, 2yyn, 2z1b, 2z5j, 2z8h, 2z9s, 2zf8, 3a0f, 3b3l, 3b43, 3ceq, 3ebm, 3fte, 3hpm, 3hxq, 3i7k, 3ikk, 3pin, 3psi, 3q2c, 3rg6, 3tn9, 3uj1, 3uw0, 3v6y, 3zor, 4az8, 4c85, 4d2g, 4dbg, 4dks, 4eig, 4fi9, 4ga7, 4i6j, 4jgy, 4jup, 4jvy, 4k0r, 4kce, 4lgs, 4lpz, 4nc2, 4nik, 4q4j, 4q5y, 4qrr, 4tkn, 4tlw, 4uy9, 4wnf, 4xfu, 4yzy, 4zyh, 5c4r, 5cca, 5cfc, 5cyu, 5d0o, 5d06, 5eta, 5eve, 5feg, 5h07, 5hpc, 5i50, 5id4, 5j6e, 5jhf, 5k23, 5k93, 5kud, 5l7c, 5lad, 5lsk, 5mqo, 5omn, 5sv7, 5tjw, 5wco.

The 64 Cambridge Structural Database (CSD) codes of the cyclic peptides that were considered in this work are as follows:

AAGAGG10, AAGGAG10, ALASAR, ALPRAL10, BIHTUH, BIHXUL10, BINJIR, BUYXOI, CACNOJ10, CAHWEN, CAMVES, CEWCIQ10, CGDLLL10, CGLEGL, CGLPGL, CGPGAP10, CLPGDH, CYBGPP, CYHEXG, DASXIE, DEWFEQ, DICWET, DUPKEE, DUTLAF10, DUVGOQ10, DUYTIA, EVAPUM, FIVSAE, GAJFAY, GEHKUC, GGAAGG, GICHOP, GIPKAR10, GOKXOV, JUXHAL, KARPIE, KIVDIC, LENKIY, LETHIE, LETPIM, LEYCAV, NIWHEH, NUCZUH, NUWNEY, PAPGAP, PAPRVA, POWWEE, PROGLY20, RUQVAB, SAFVOM, SEFTIG, SOWGOA, TALVAD, UBADEJ, UNONES, UZUKUW, VAWTAQ, WUYGII, YEXJIV, YOMNOE, ZAJPAB, ZEHDEV, ZOHMIS, ZUKRAY.

The 39 bioactive peptide sequences in terms of their one letter amino acid codes as well as the associated bio-functionality as reported in literature[1] are as follows:

Peptide sequence	**Bio-functionality[1]**
RGD, RGDS, KGD, PHSRN, REDV, YIGSR, IKVAV, PDSGR, DGEA, KRSR, GFPGER	Cell adhesion peptides
$(GAGA)_2$, $(GPP)_2$	Structural peptides
CNGRC, CGKRK, CRGDK, CREKA	Anti-tumor peptides
RRWWRF, FRWWHR, KLAK	Anti-microbial peptides
VYIHPF, WMNF	Peptide hormone
GPQGIAG, APGL, VRN	Tissue-engineering application
GNNQQNY, VQIVYK, NFGAIL, KLVFF, KLVFFAE, LPFFD, FEFEFKEK	Model amyloid peptides
YGGFM, YGGFL, YPWF, YPFF	Neuropeptides
RLNVY, RLGVY	Immune-related peptides
HHHHHH	Protein tags

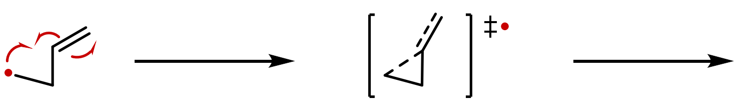
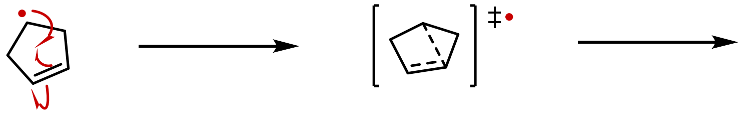
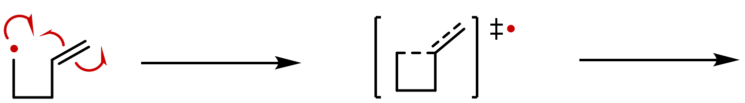
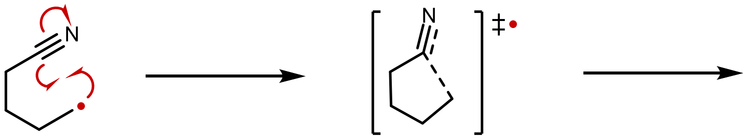
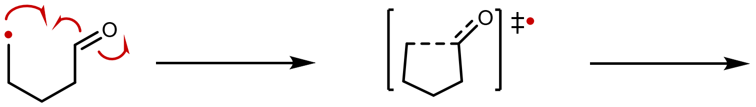
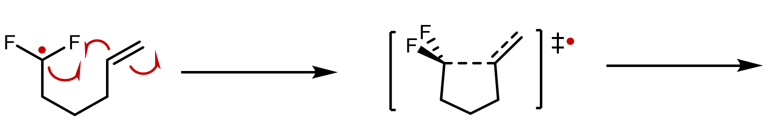
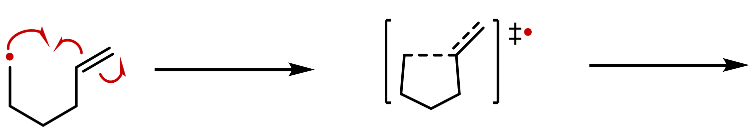
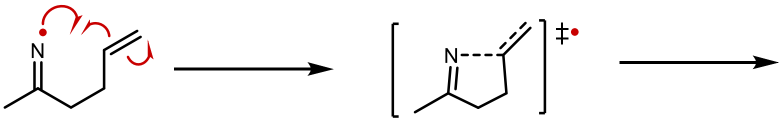
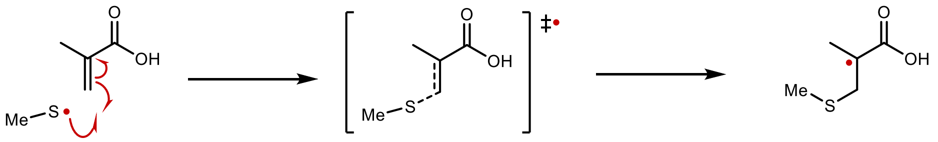
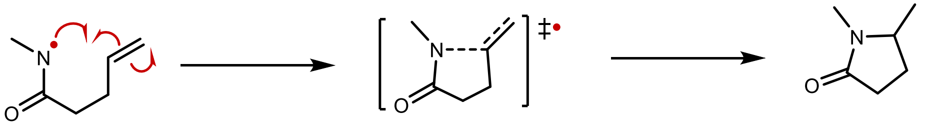
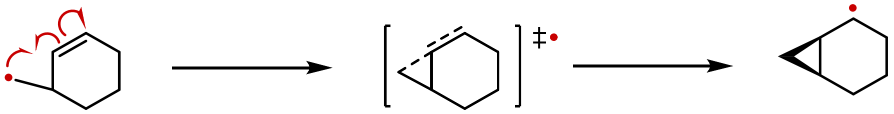
[1] Hamley, I. W. Small bioactive peptides for biomaterials design and therapeutics. *Chem. Rev.* **117,** 14015–14041 (2017).

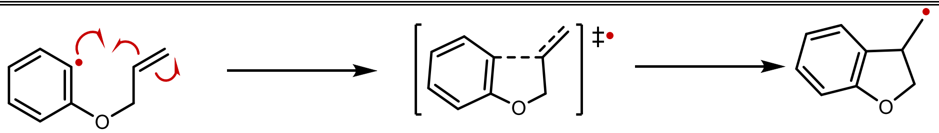
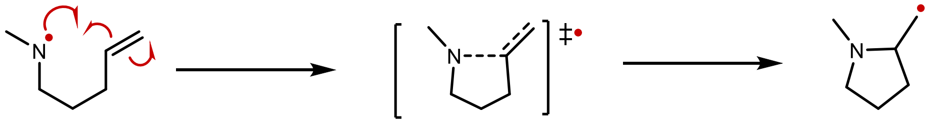
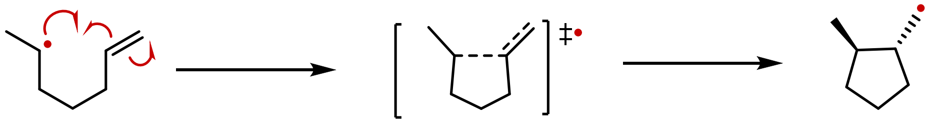
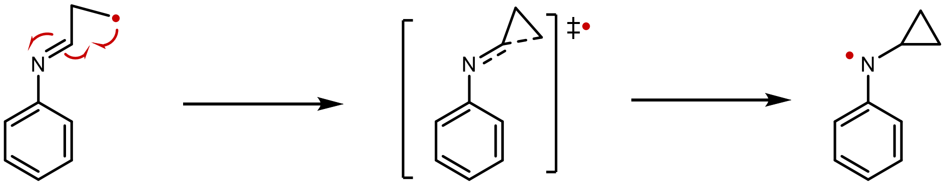
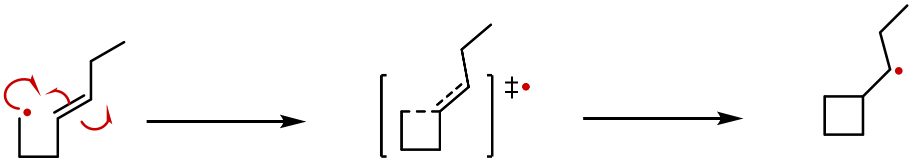
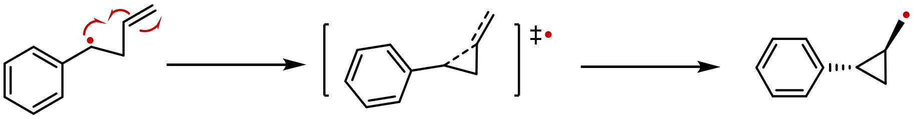
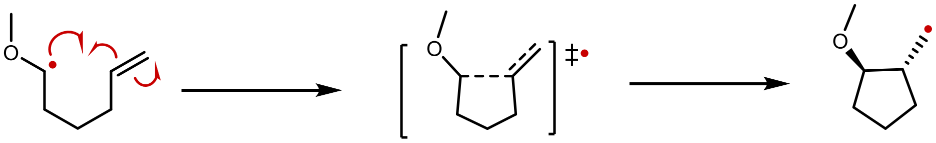
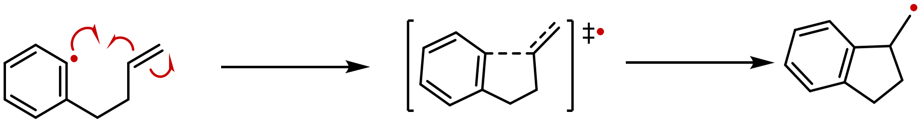
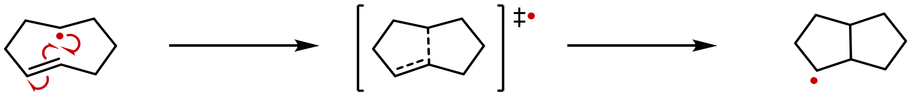
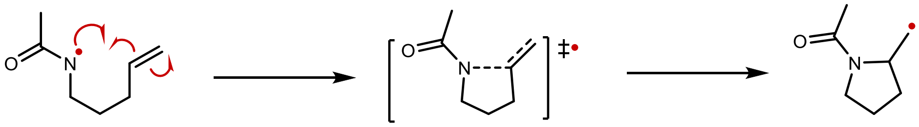
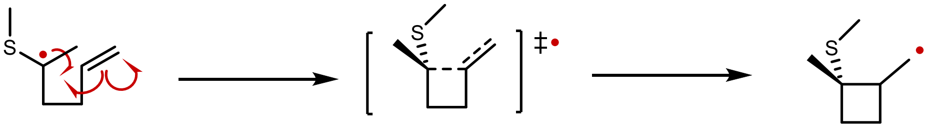
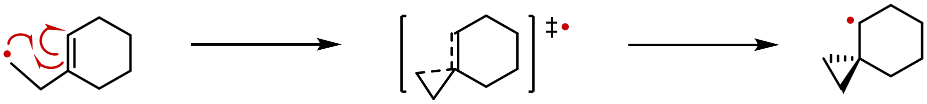
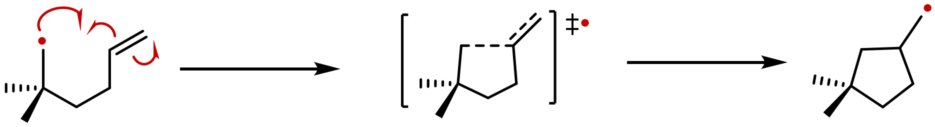
Appendix 3

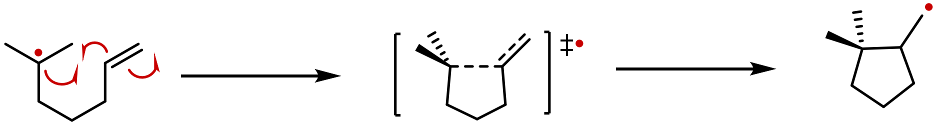
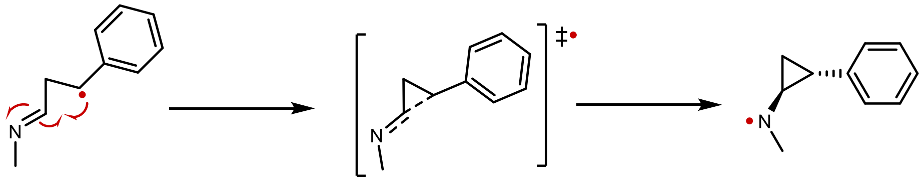
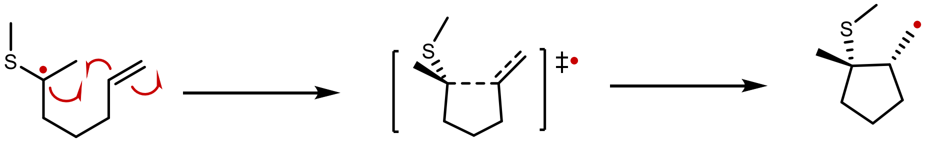
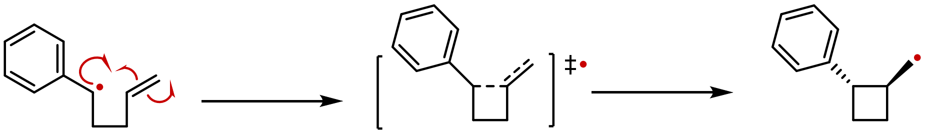
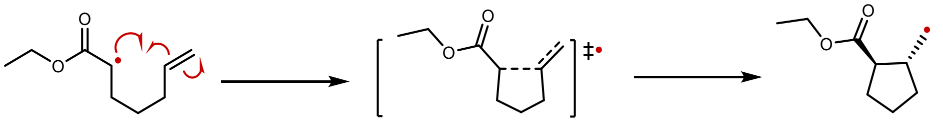
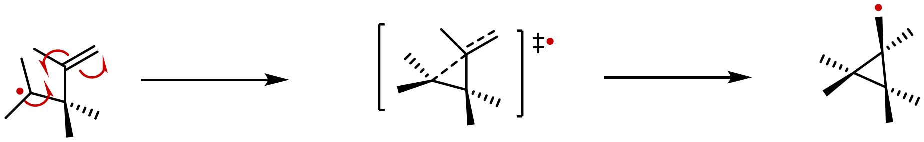
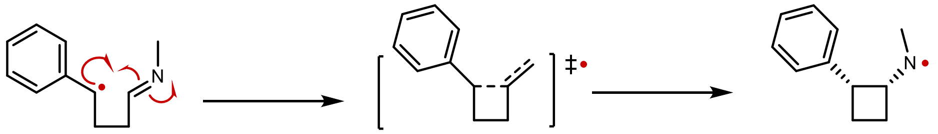
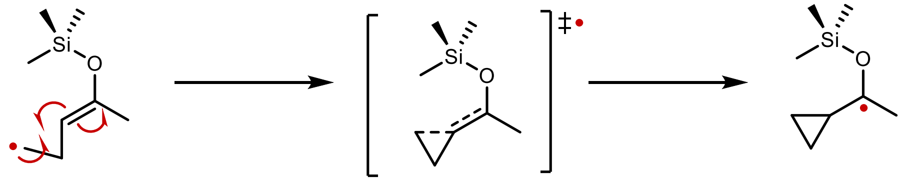
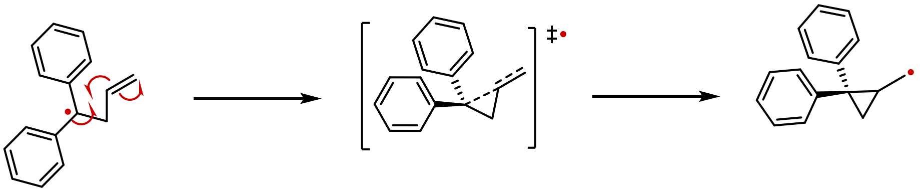
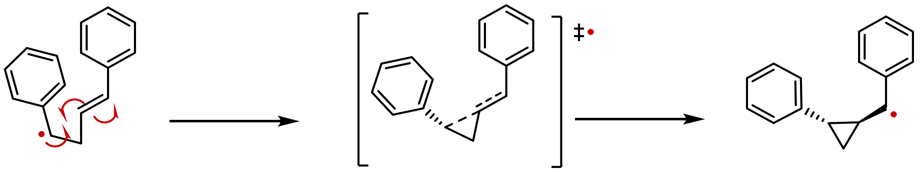
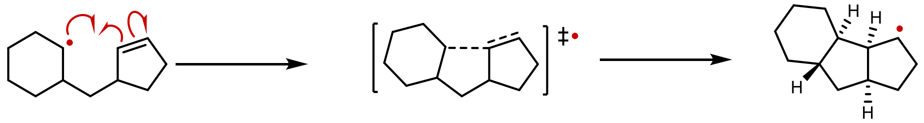
Supporting Information to Chapter 6

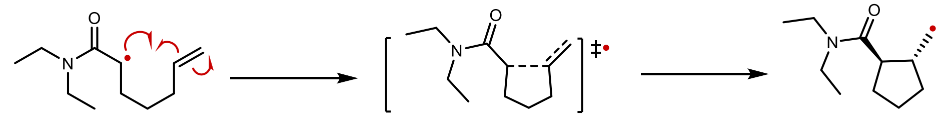
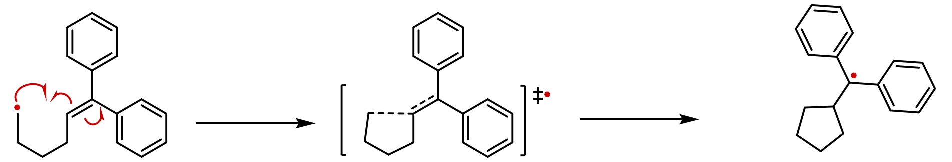
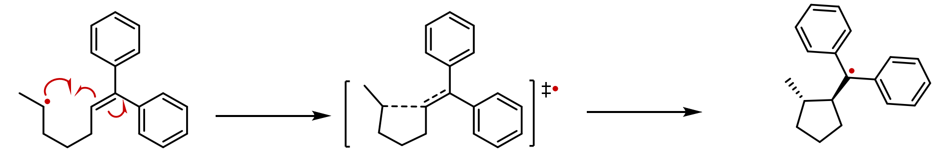
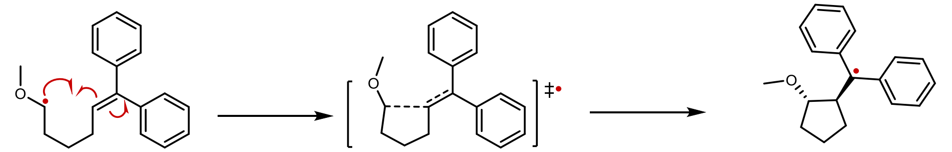
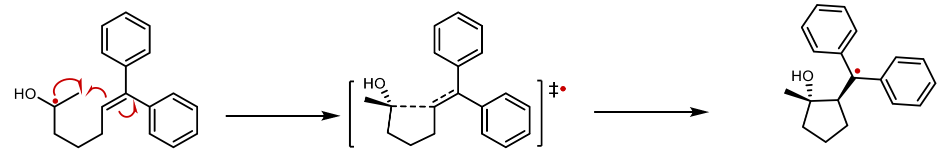
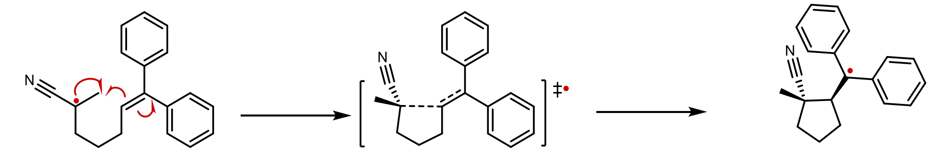
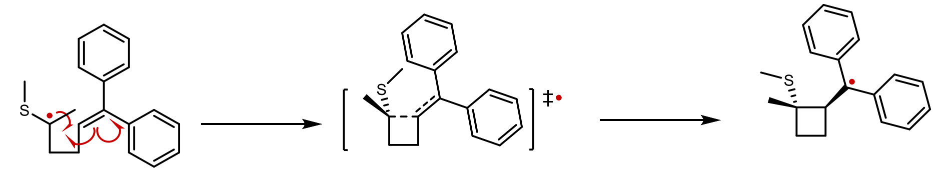
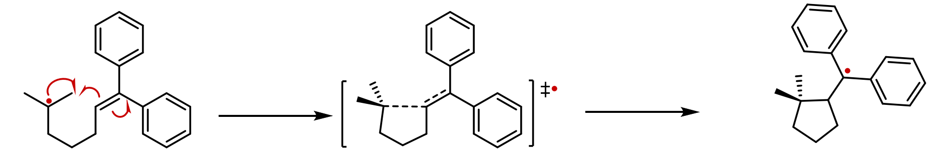
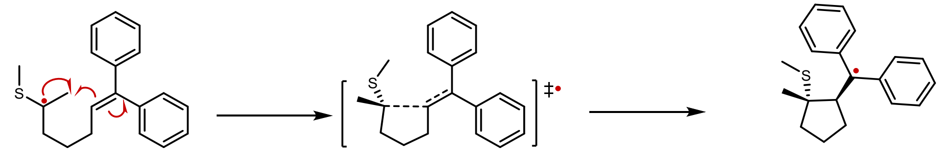
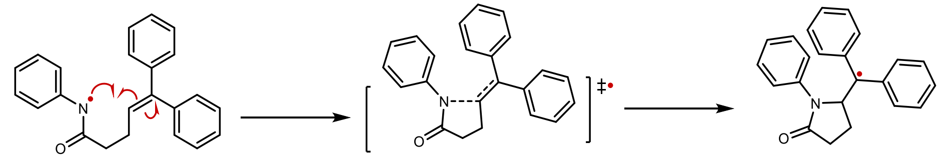
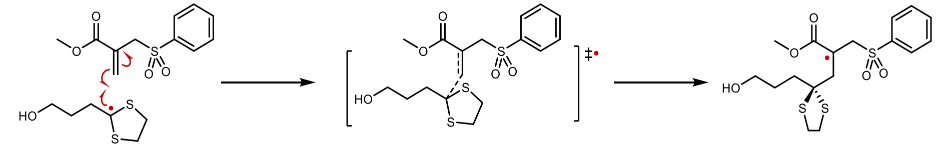
Section S1. Reactions in the BH9 data set

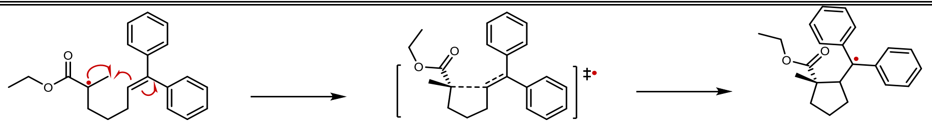
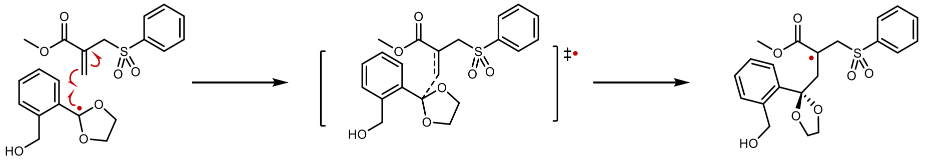
I. Radical rearrangement and addition

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
1		11.74	9.01	2.74
2		38.51	6.56	31.94
3		17.10	15.18	1.92
4		10.80	20.54	-9.74
5		8.56	10.84	-2.28
6		6.92	26.65	-19.73
7		7.94	24.57	-16.63
8		13.67	20.45	-6.77
9		-0.79	15.79	-16.58
10		7.64	31.53	-23.89
11		10.88	8.30	2.58

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
12		3.37	36.53	-33.16
13		14.83	20.64	-5.81
14		7.96	23.95	-15.99
15		9.35	11.72	-2.37
16		17.04	15.20	1.85
17		16.96	3.87	13.09
18		7.76	19.98	-12.22
19		4.25	36.12	-31.88
20		5.78	30.88	-25.10
21		5.42	25.97	-20.55
22		16.38	9.26	7.13
23		12.96	9.63	3.33
24		7.11	24.94	-17.82

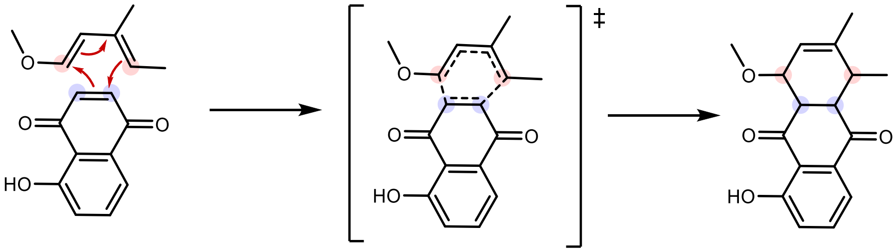
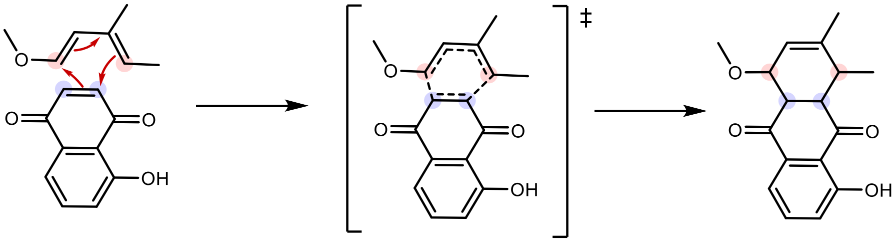
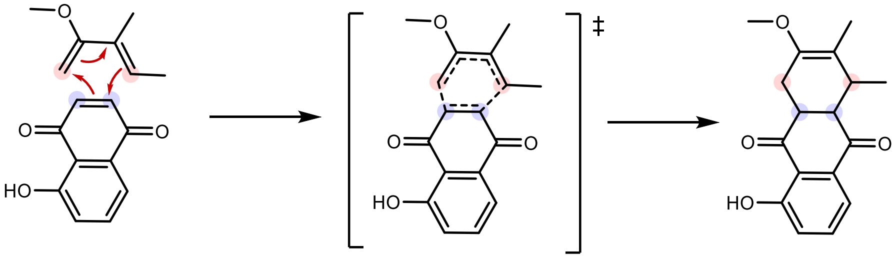
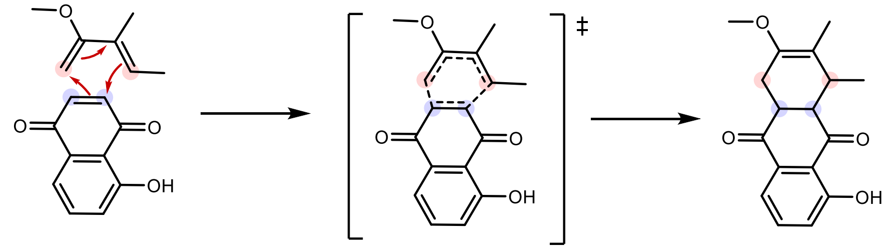
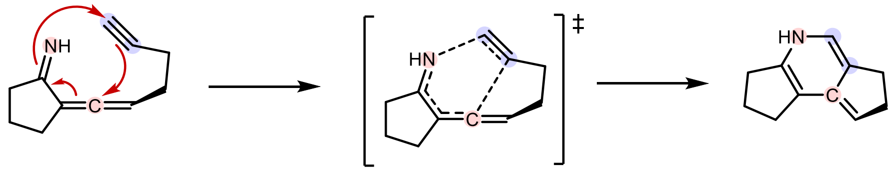
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
25		6.75	21.05	-14.30
26		18.76	2.40	16.36
27		9.78	18.83	-9.05
28		22.46	9.64	12.82
29		7.92	18.08	-10.16
30		14.29	6.03	8.26
31		24.18	7.64	16.54
32		11.47	10.55	0.92
33		20.07	3.03	17.04
34		13.02	8.08	4.94
35		9.68	27.80	-18.12

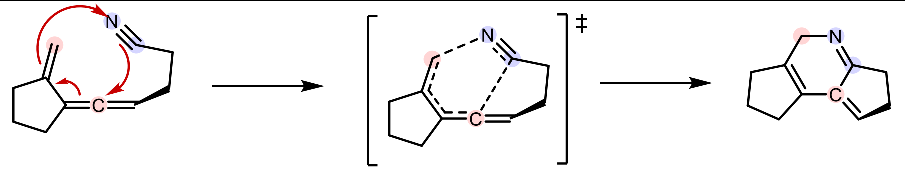
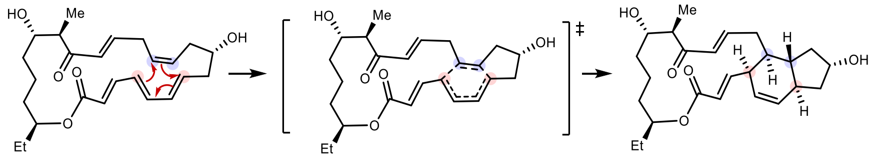
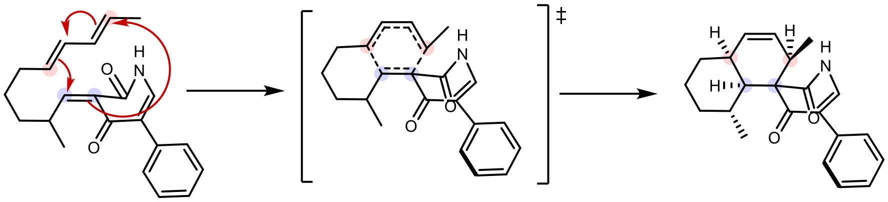
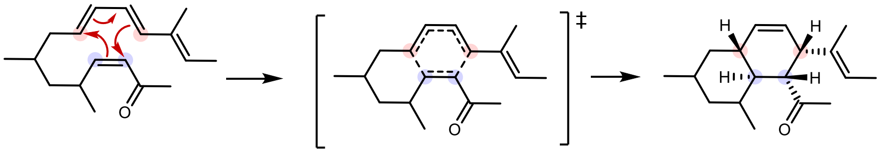
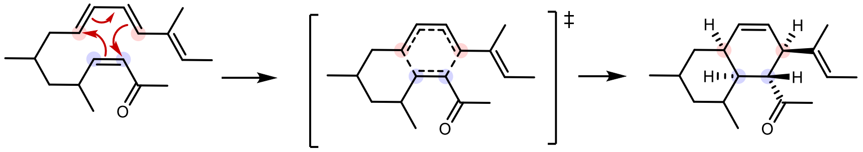
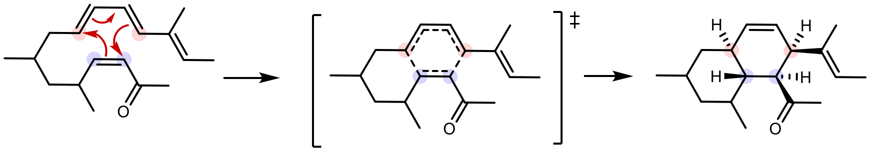
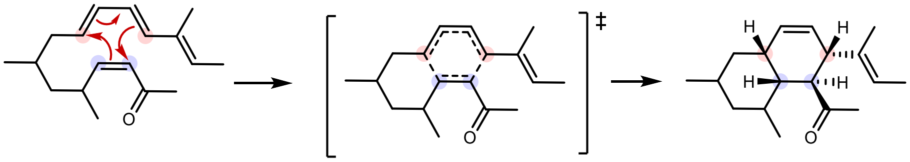
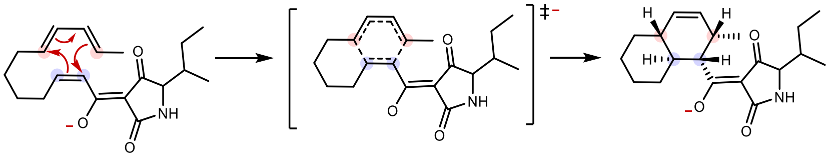
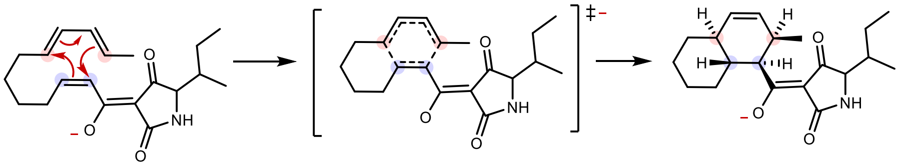
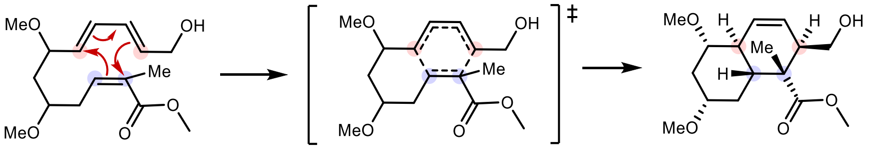
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
36		7.82	18.49	-10.67
37		6.15	31.91	-25.76
38		6.29	30.93	-24.64
39		5.30	27.45	-22.15
40		5.71	24.84	-19.13
41		7.58	-15.38	22.96
42		13.31	13.52	-0.22
43		5.45	25.56	-20.11
44		6.10	22.21	-16.11
45		-26.78	-5.68	-21.10
46		0.01	19.72	-19.71

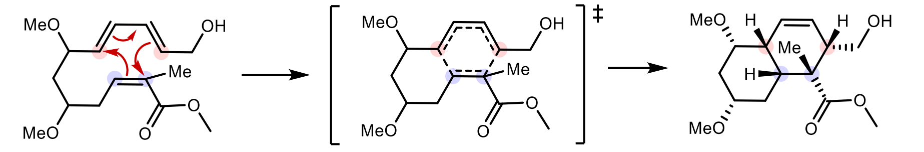
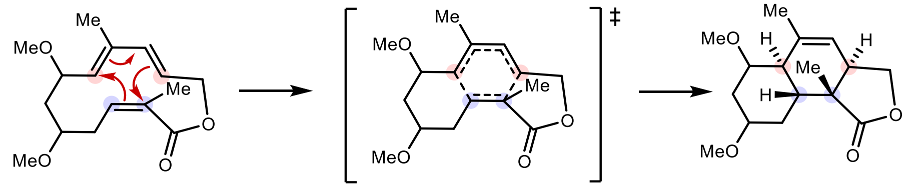
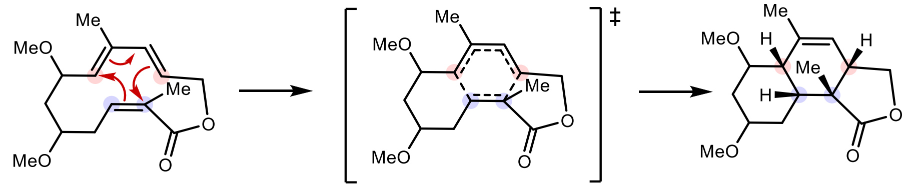
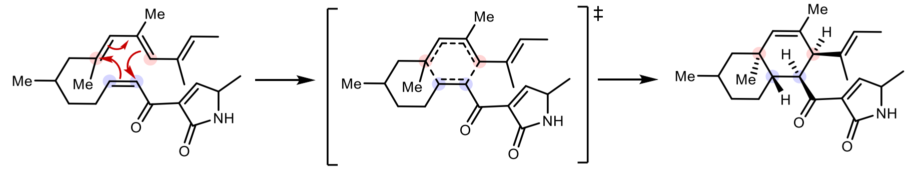
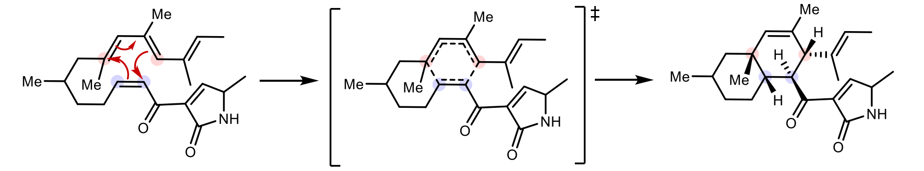
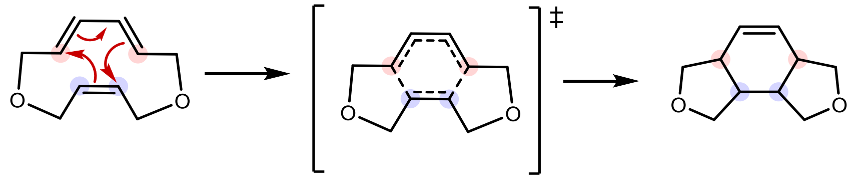
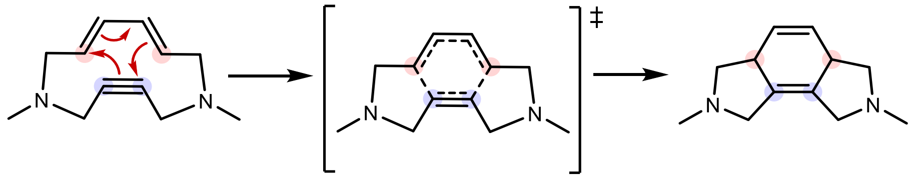
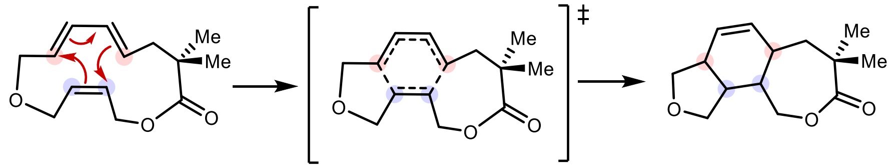
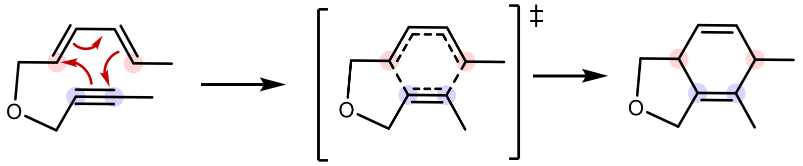
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
47		7.47	58.28	-50.81
48		0.37	22.46	-22.09

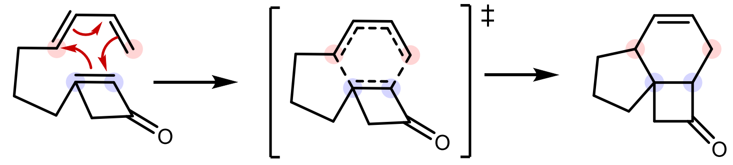
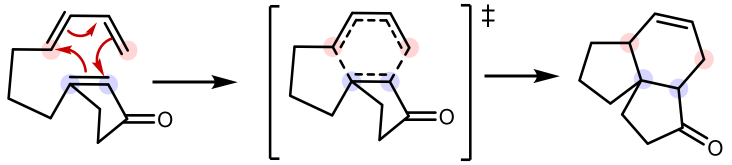
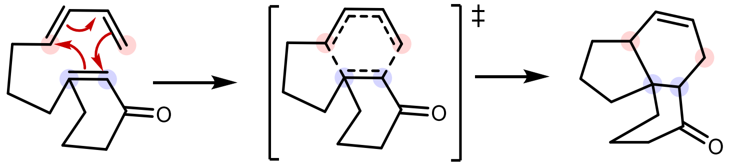
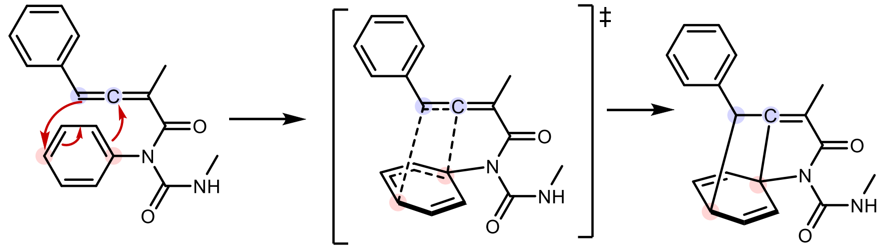
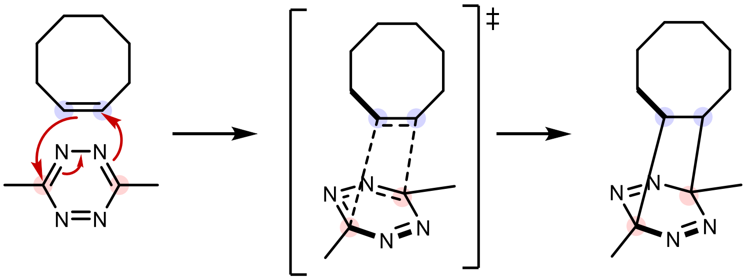
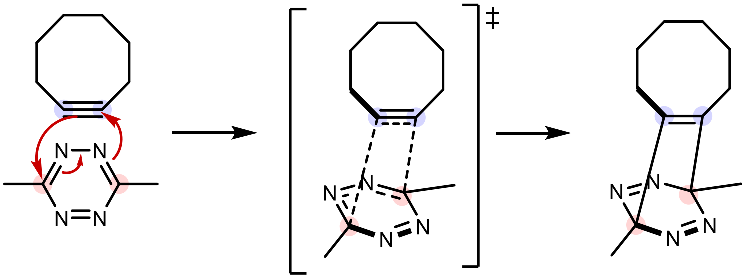
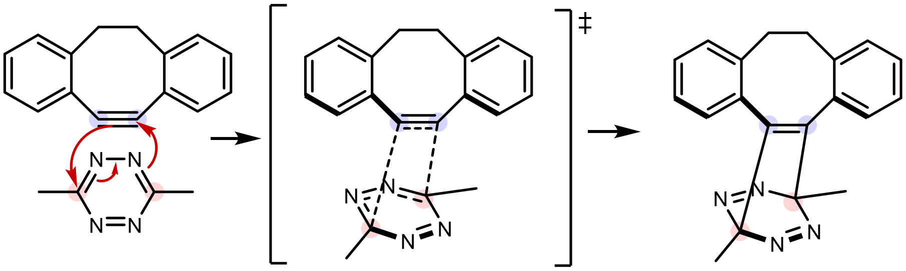
[a] units are in kcal/mol.

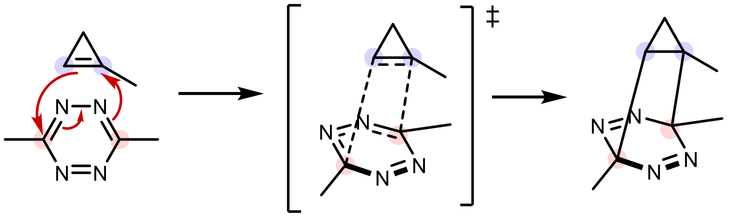
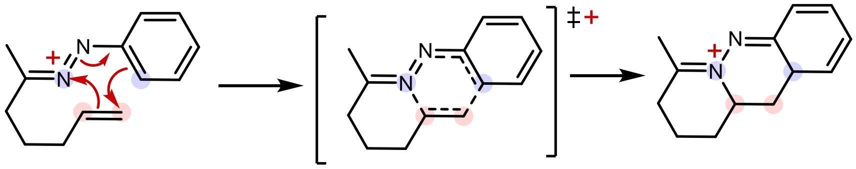
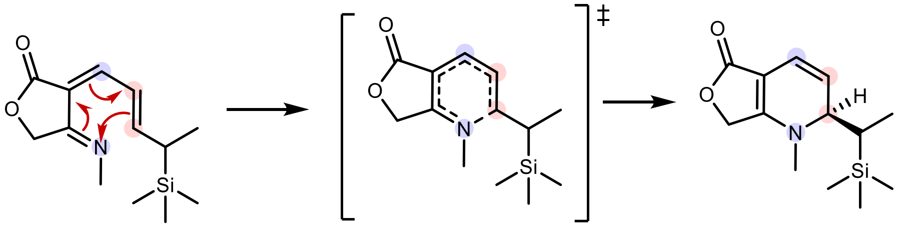
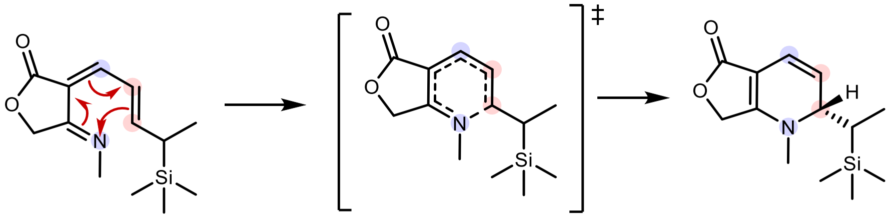
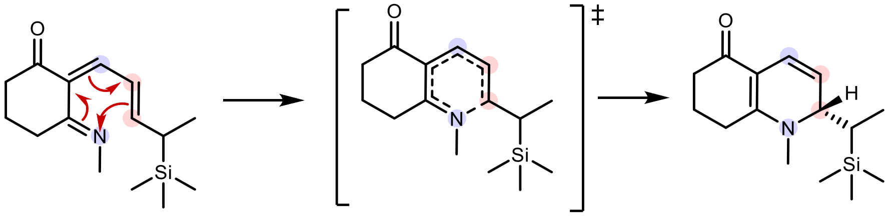
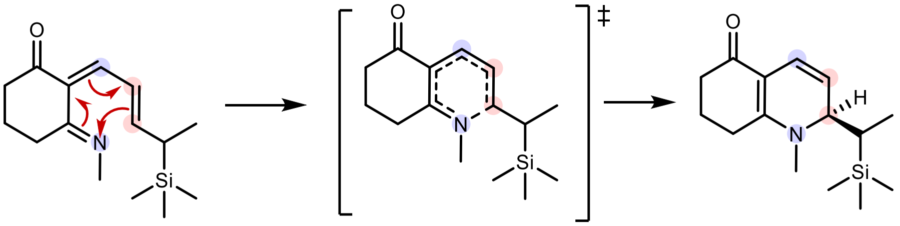
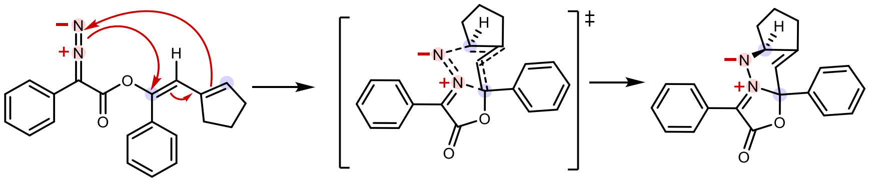
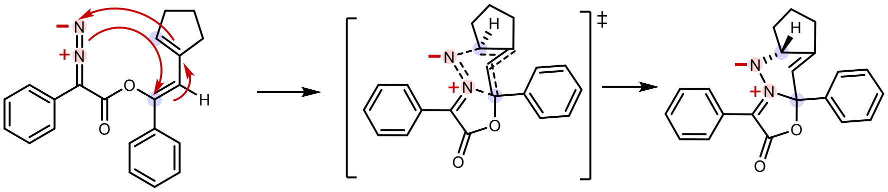
II. Pericyclic

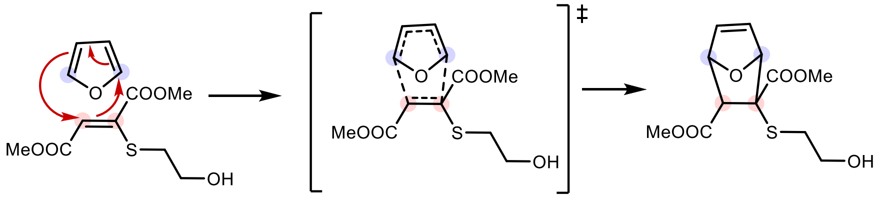
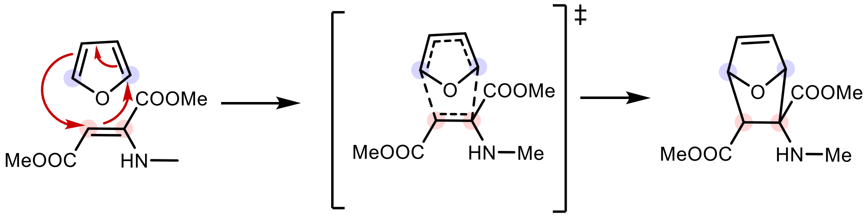
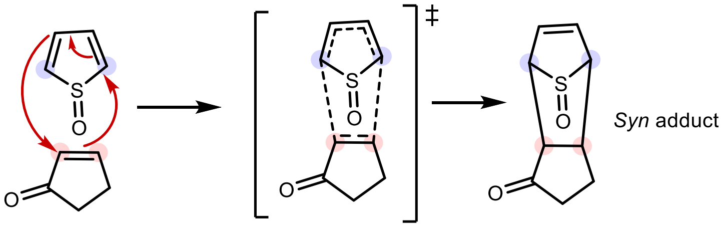
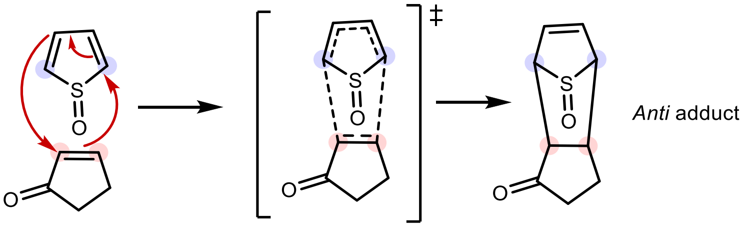
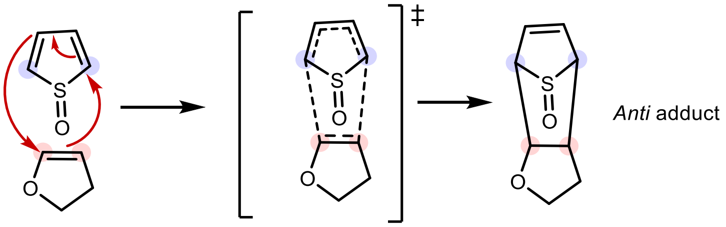
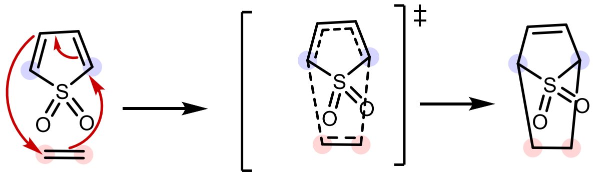
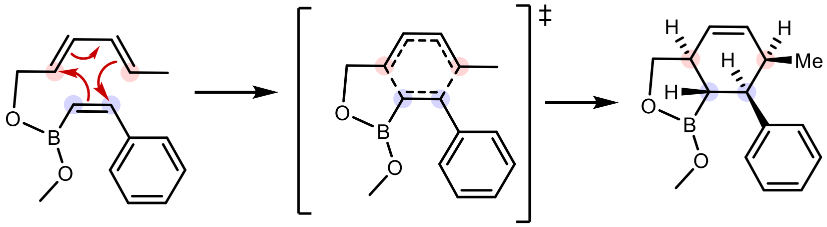
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
1		8.10	43.77	-35.66
2		8.85	43.49	-34.63
3		11.45	47.83	-36.38
4		10.15	47.06	-36.92
5		26.57	93.72	-67.15

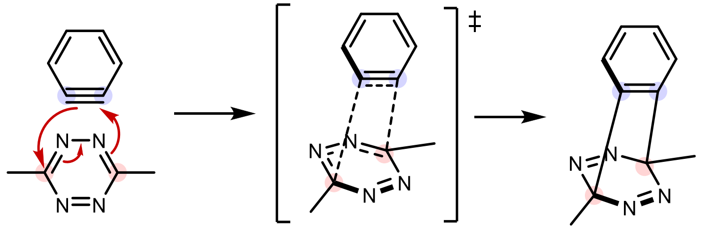
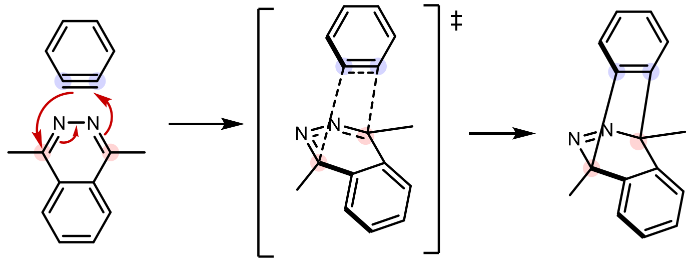
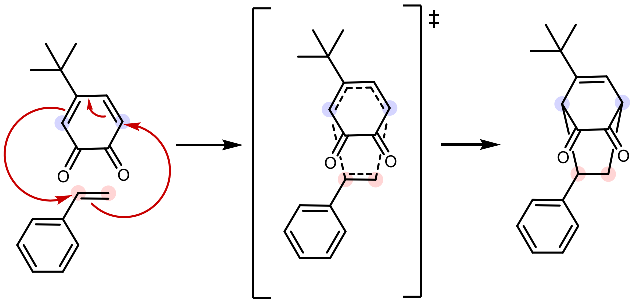
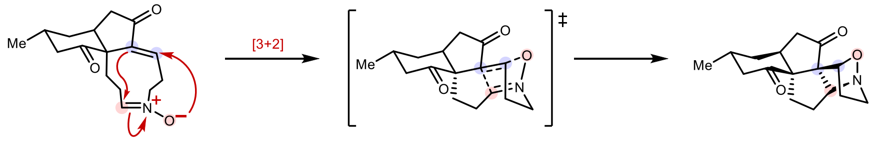
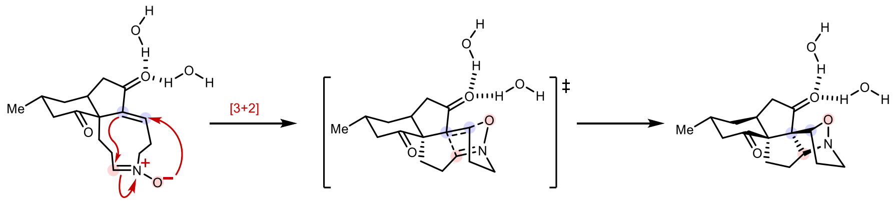
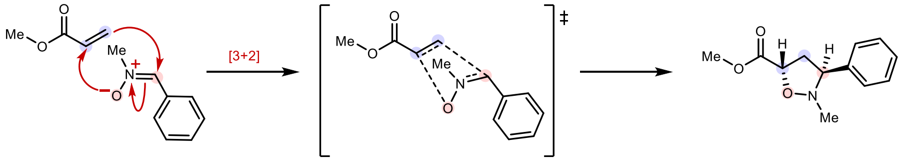
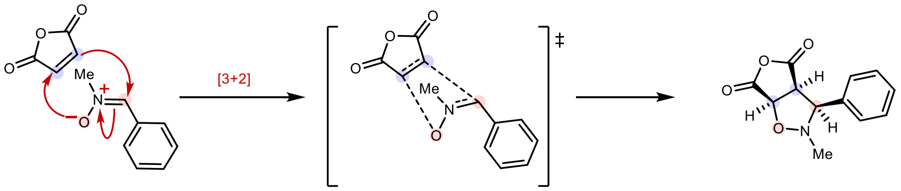
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
6		26.77	62.05	-35.27
7		19.07	39.40	-20.32
8		12.20	36.77	-24.57
9		17.83	47.85	-30.02
10		19.75	50.92	-31.17
11		23.08	45.70	-22.62
12		24.51	44.87	-20.36
13		24.90	60.92	-36.03
14		21.97	60.48	-38.51
15		25.29	60.68	-35.39

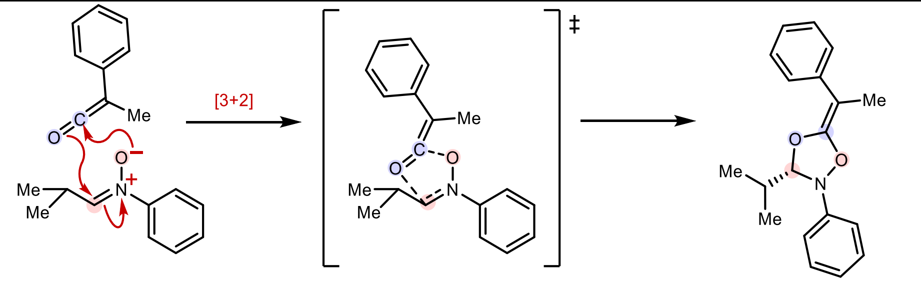
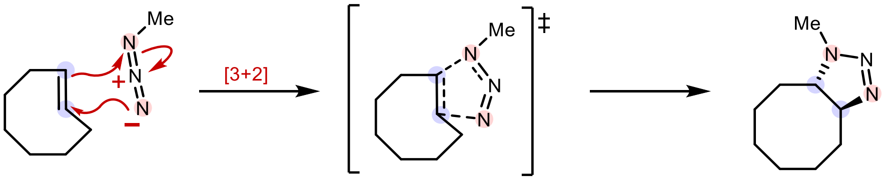
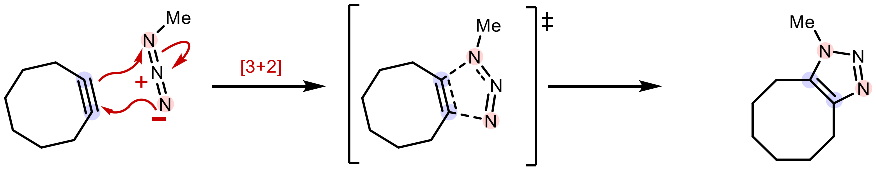
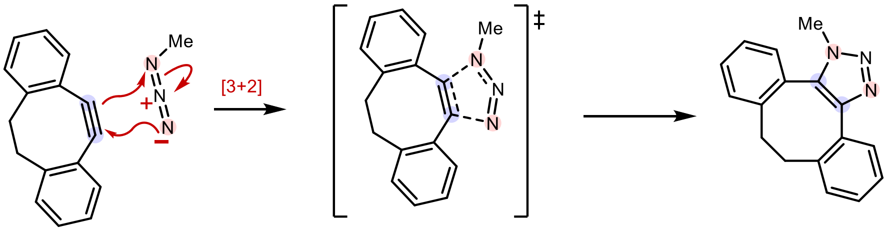
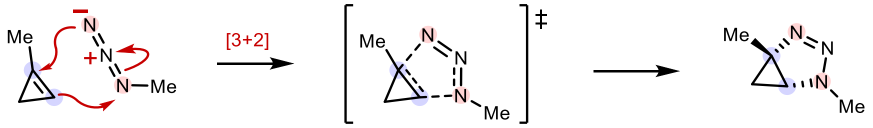
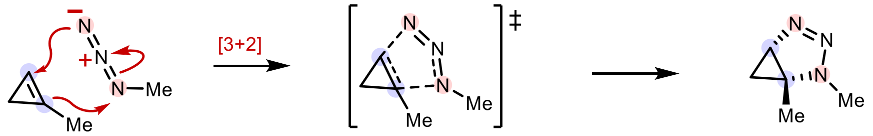
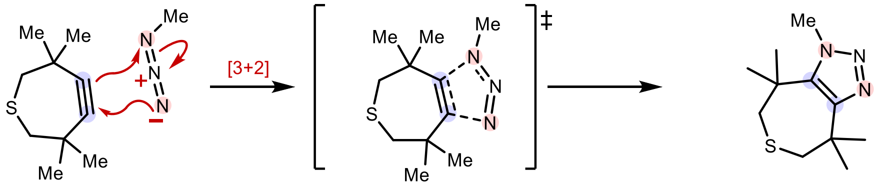
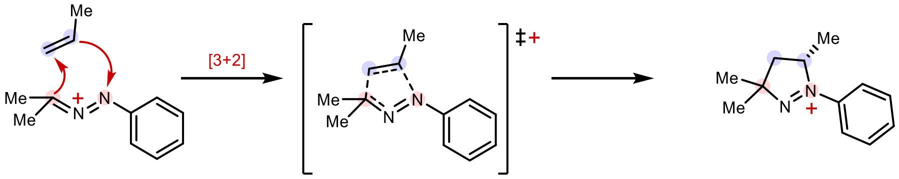
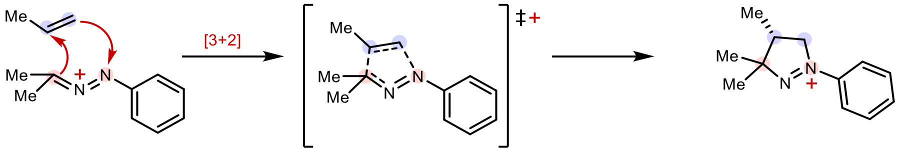
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
16		23.43	53.32	-29.89
17		15.27	59.46	-44.18
18		19.86	70.40	-50.54
19		22.04	59.33	-37.29
20		19.10	56.89	-37.79
21		12.40	62.14	-49.74
22		8.83	68.85	-60.01
23		24.81	57.02	-32.21
24		26.62	73.95	-47.33

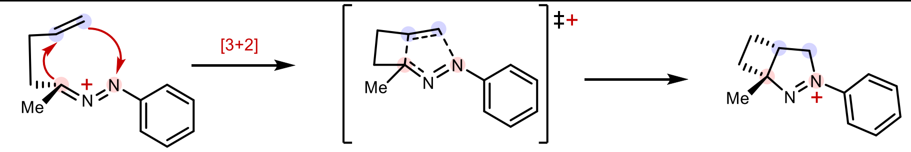
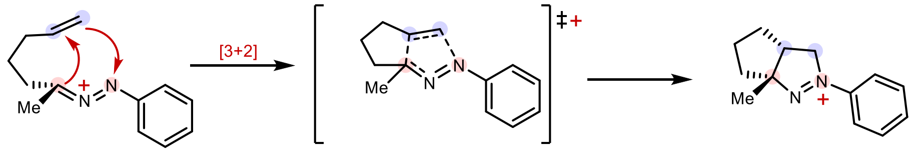
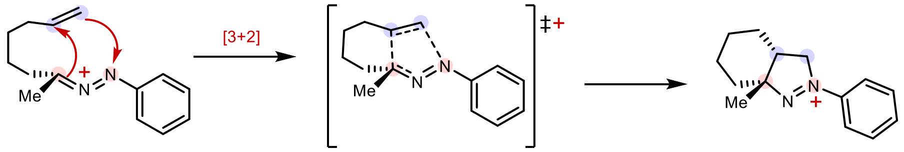
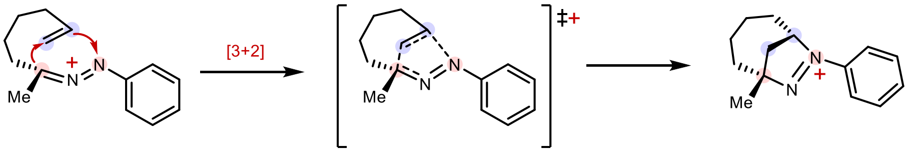
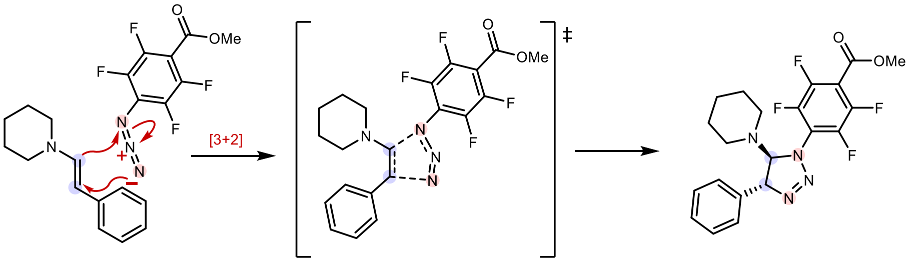
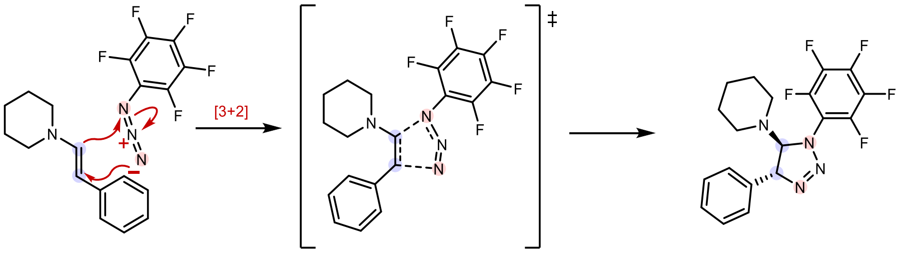
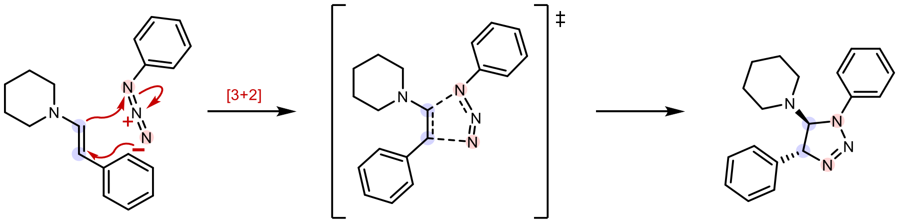
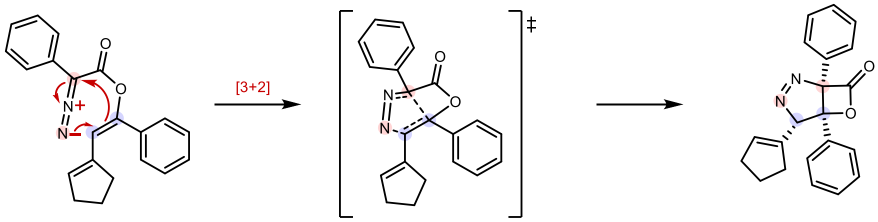
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
25		23.55	54.52	-30.97
26		27.83	47.79	-19.96
27		29.60	50.22	-20.62
28	 C O N NH O ‡ C O N NH O C O N NH O	21.69	38.32	-16.63
29		1.65	35.57	-33.92
30		7.98	55.02	-47.04
31		15.12	36.68	-21.56

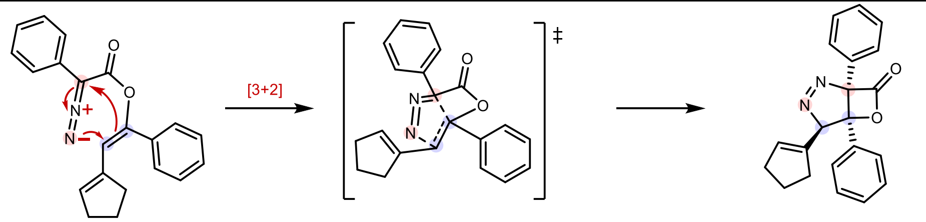
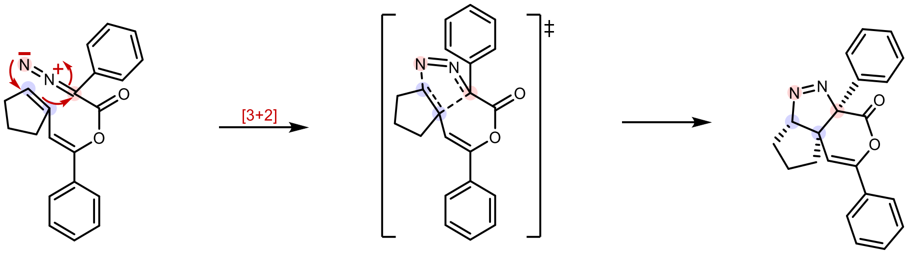
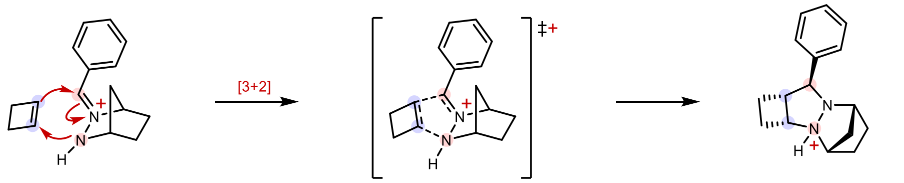
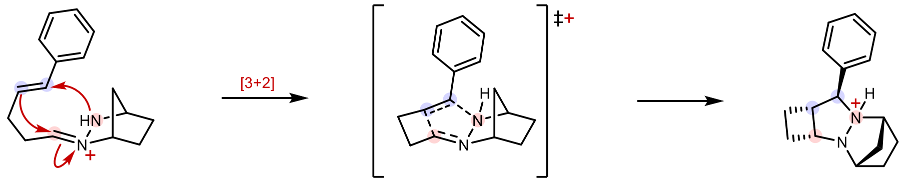
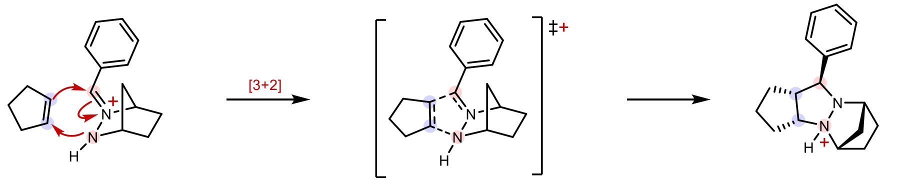
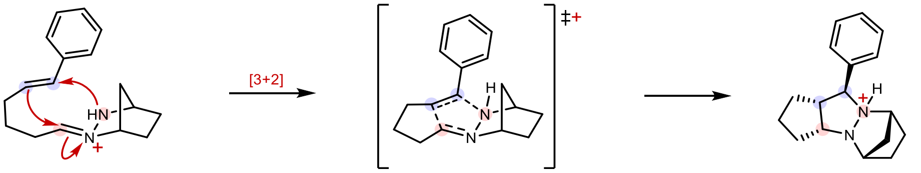
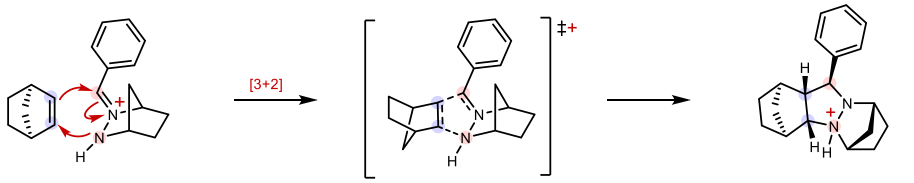
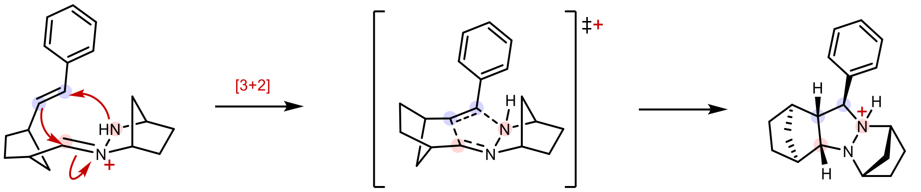
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
32		6.61	54.79	-48.18
33		11.62	53.25	-41.63
34		18.28	30.48	-12.21
35		16.53	29.81	-13.28
36		15.49	33.65	-18.16
37		17.00	34.43	-17.43
38		22.73	36.09	-13.36
39		37.42	39.32	-1.90

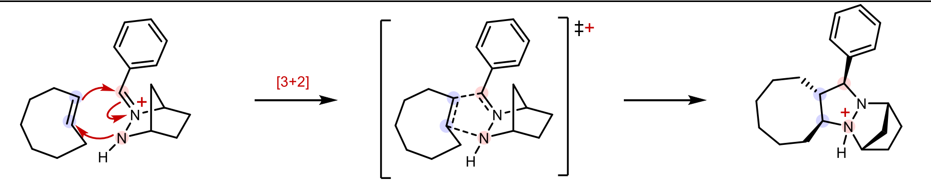
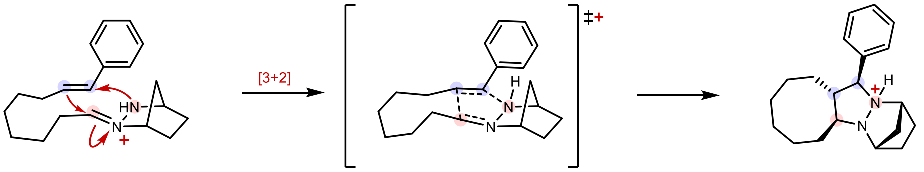
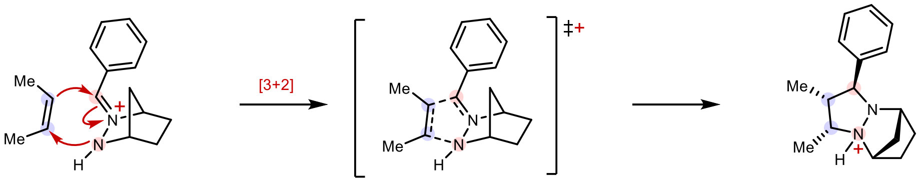
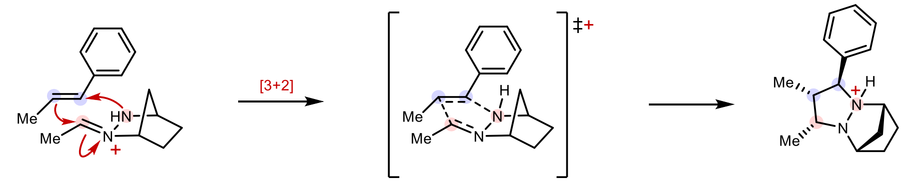
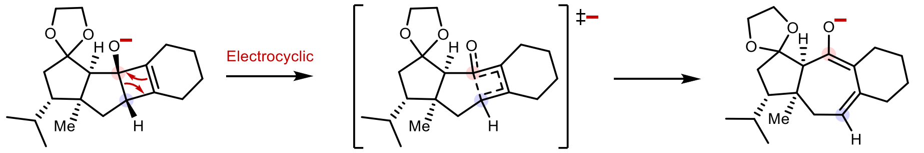
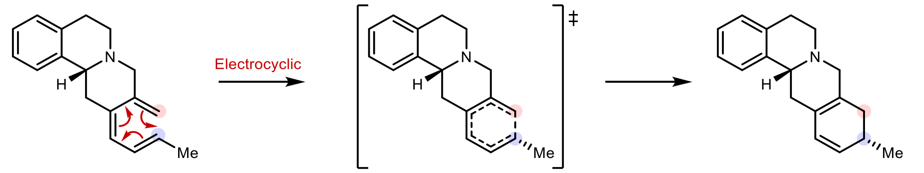
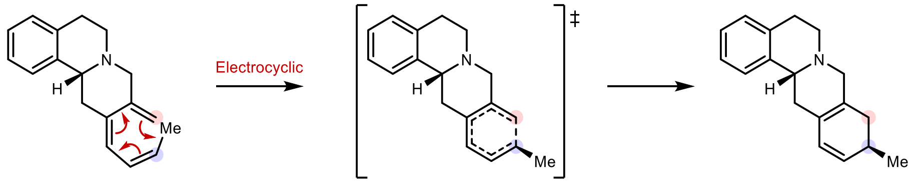
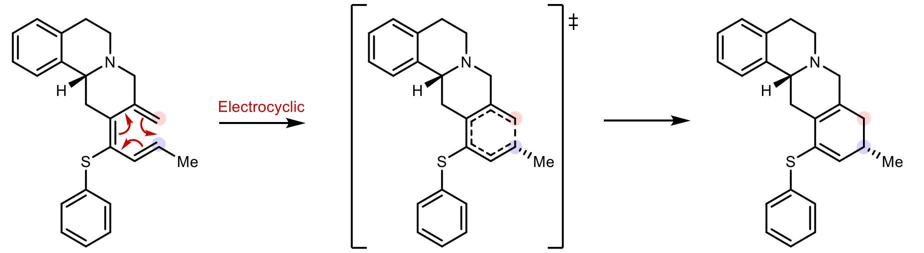
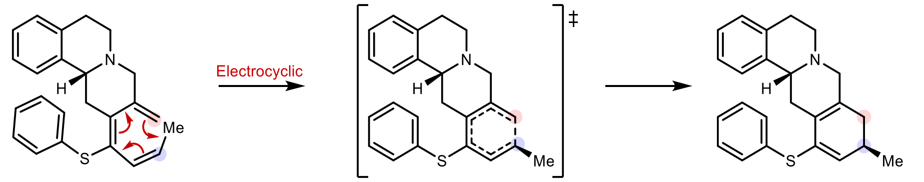
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
40		14.26	27.77	-13.51
41		28.06	31.78	-3.72
42		8.58	47.74	-39.15
43		16.35	48.76	-32.41
44		16.64	53.18	-36.54
45		16.98	59.46	-42.48
46		26.48	51.36	-24.88

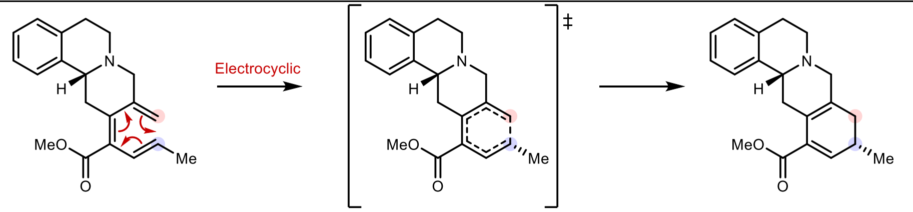
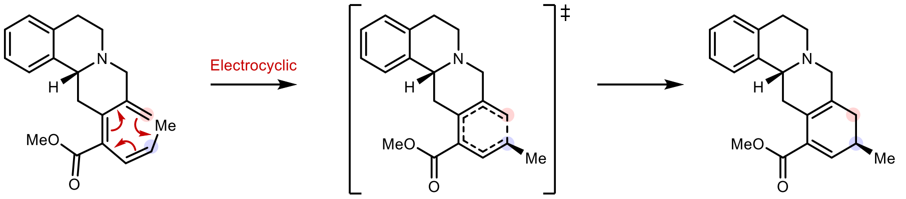
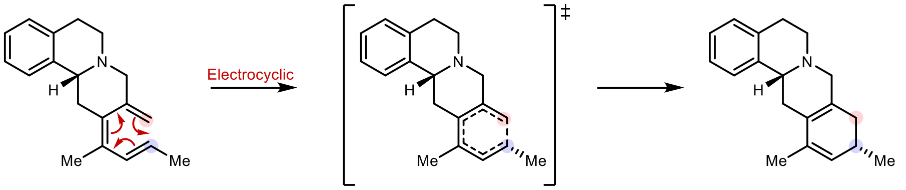
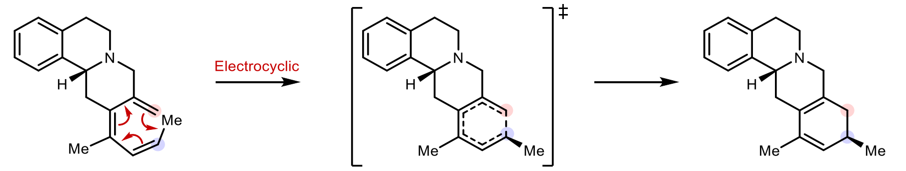
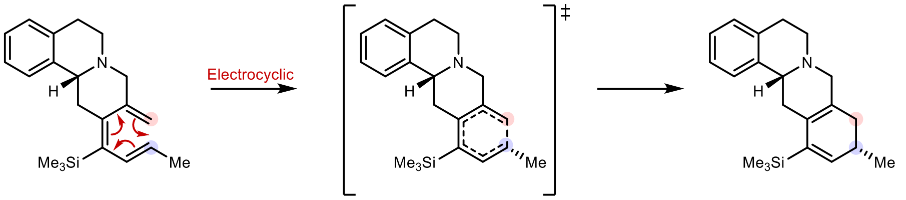
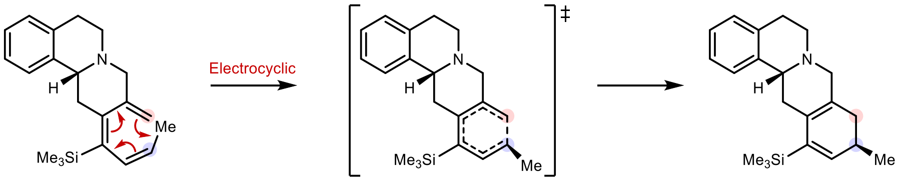
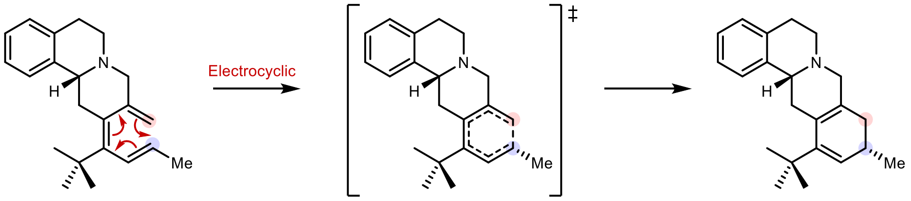
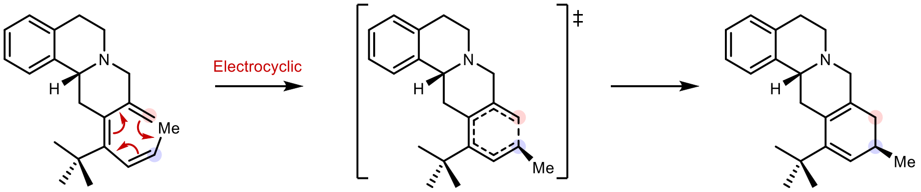
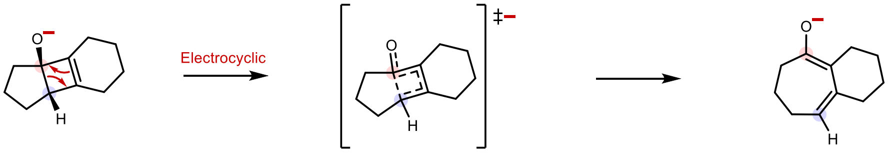
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
47		-1.34	88.51	-89.85
48		1.98	86.44	-84.47
49		13.95	47.20	-33.24
50		18.44	35.48	-17.04
51		-1.95	24.26	-26.21
52		12.39	36.11	-23.72
53		10.45	35.89	-25.44

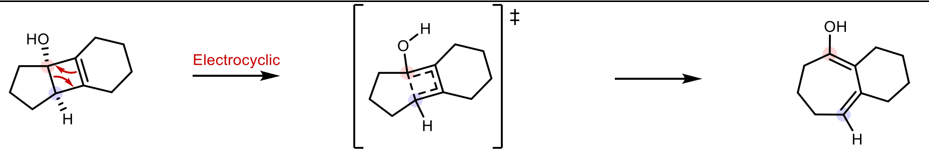
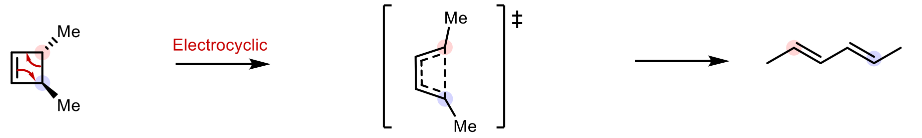
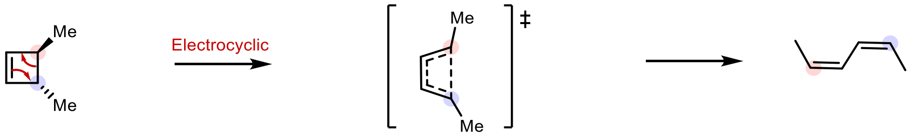
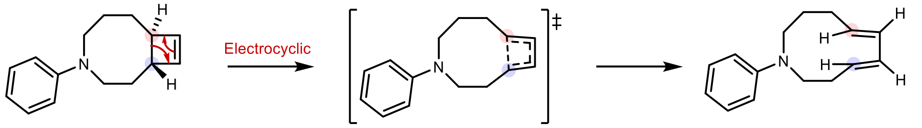
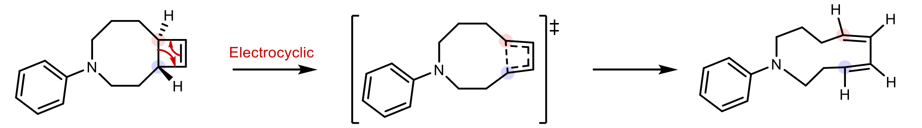
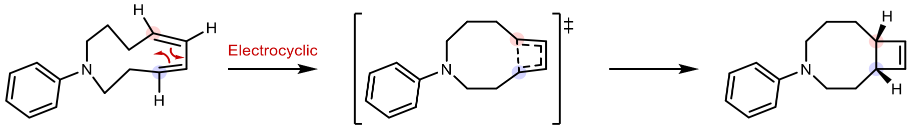
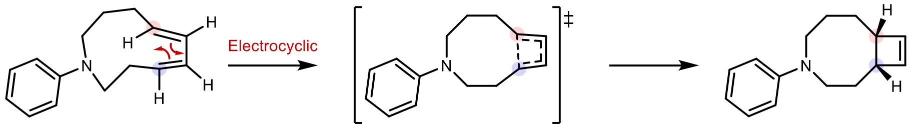
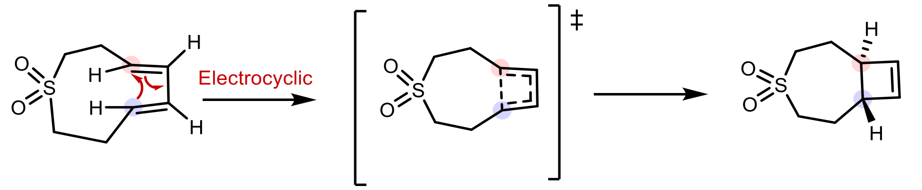
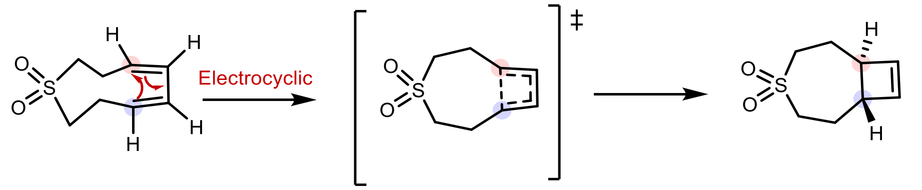
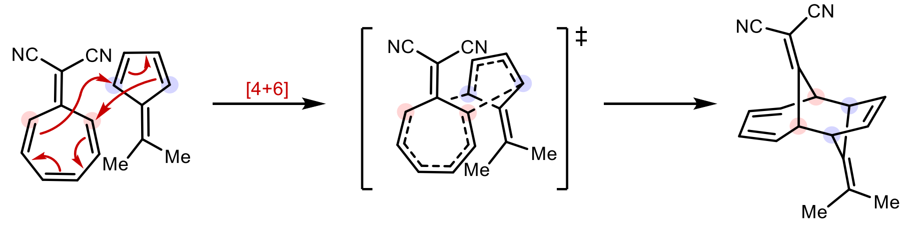
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
54		-0.12	26.38	-26.50
55		12.65	52.78	-40.12
56		11.60	91.82	-80.21
57		8.10	84.96	-76.87
58		15.43	68.29	-52.86
59		15.30	67.99	-52.69
60		5.93	86.94	-81.00
61		0.41	56.60	-56.19
62		3.87	58.42	-54.55

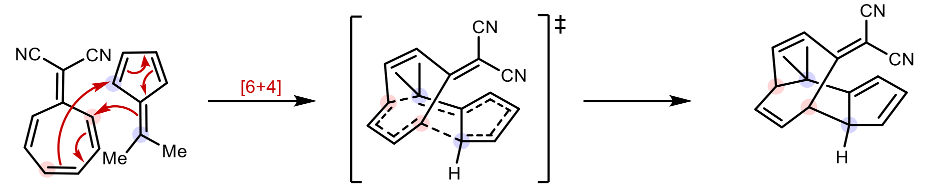
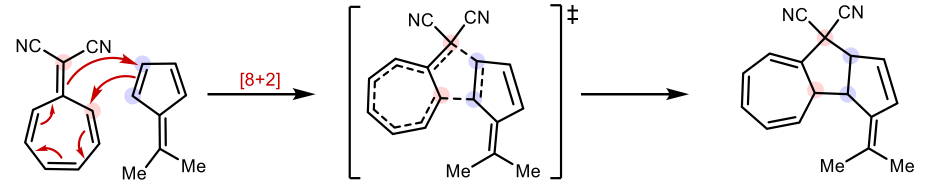
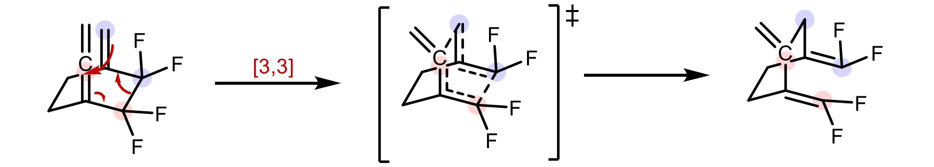
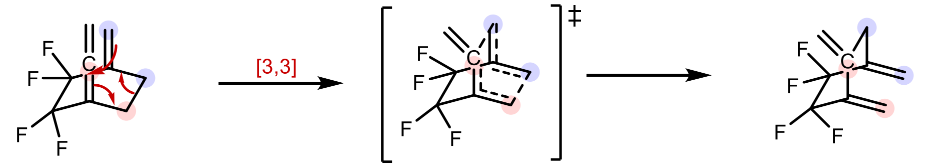
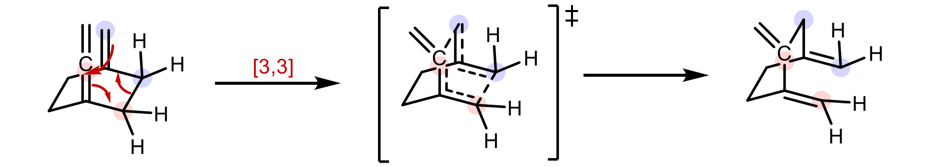
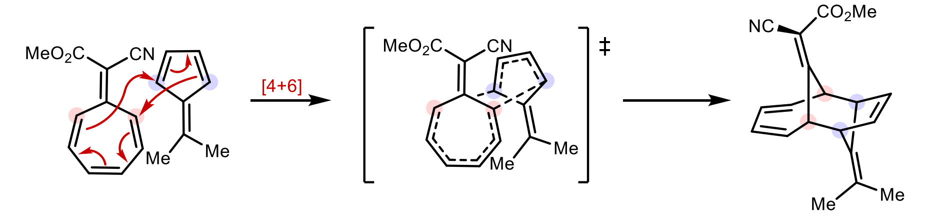
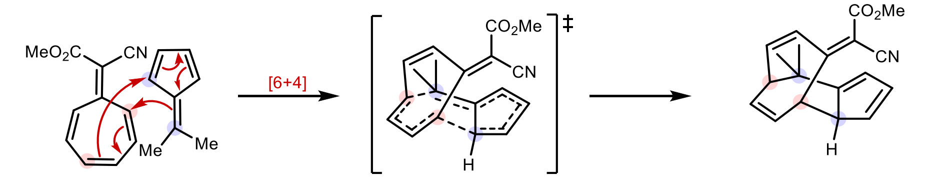
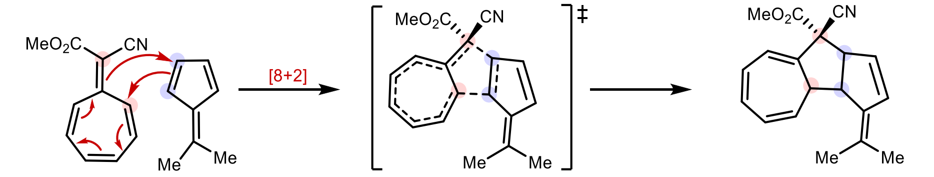
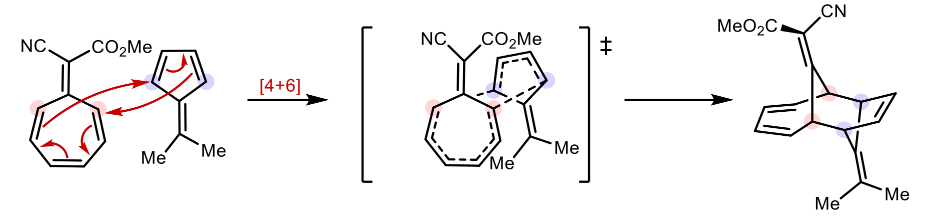
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
63		16.87	41.50	-24.63
64		5.38	50.91	-45.54
65		8.08	56.87	-48.79
66		6.79	55.77	-48.98
67		6.93	34.09	-27.16
68		8.70	35.15	-26.45
69		13.35	38.42	-25.07
70		33.30	41.80	-8.51

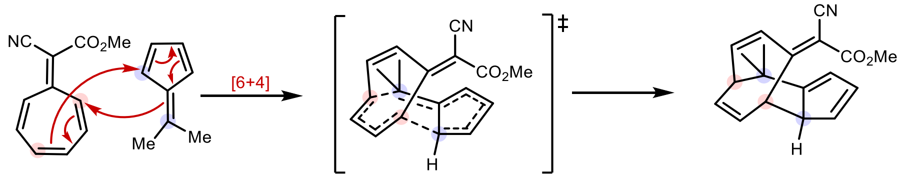
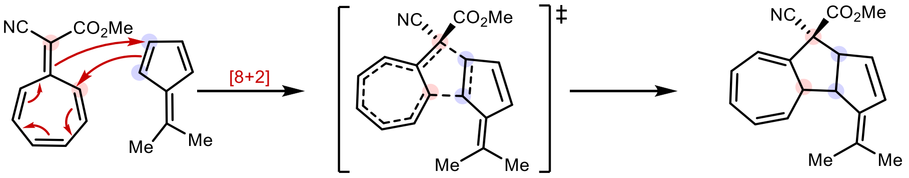
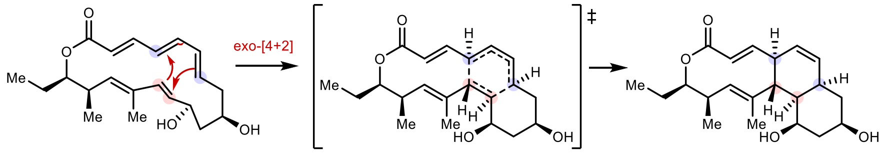
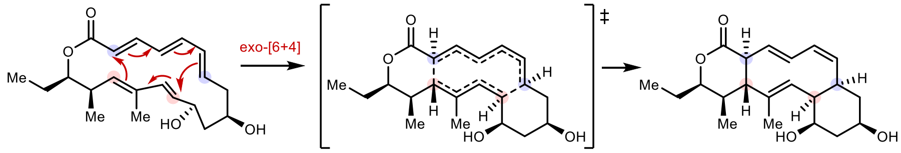
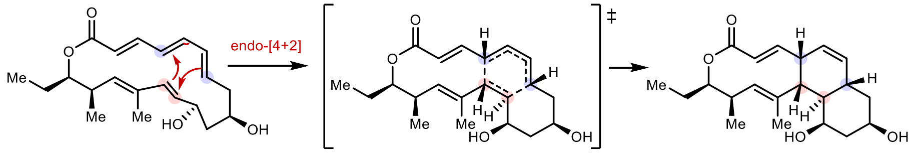
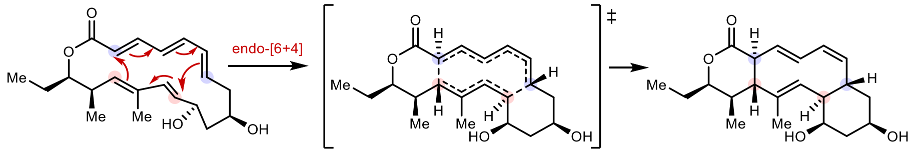
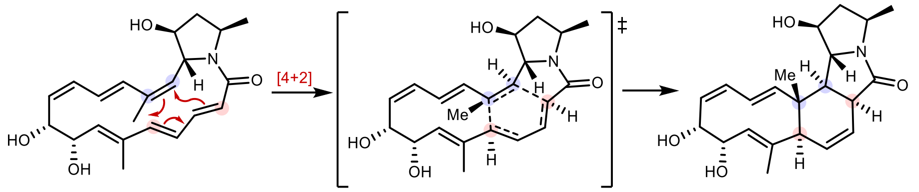
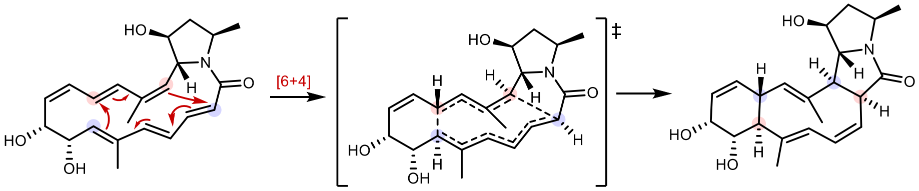
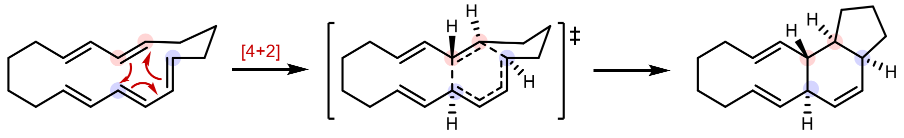
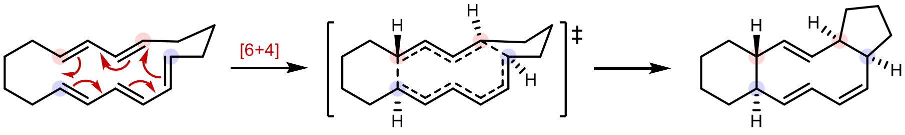
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
71		34.78	40.16	-5.38
72		26.38	51.44	-25.06
73		10.06	45.53	-35.47
74		27.86	26.23	1.63
75		9.79	41.83	-32.04
76		20.35	37.37	-17.02
77		6.99	46.28	-39.30
78		19.13	38.27	-19.14

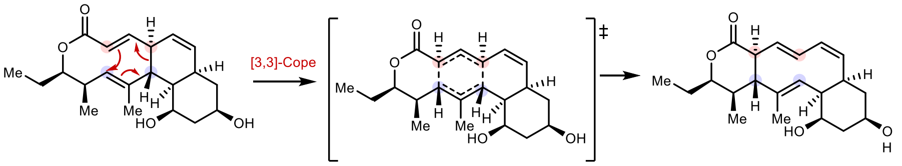
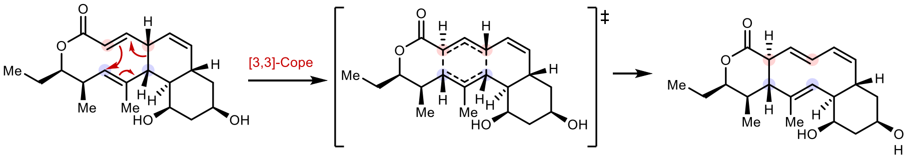
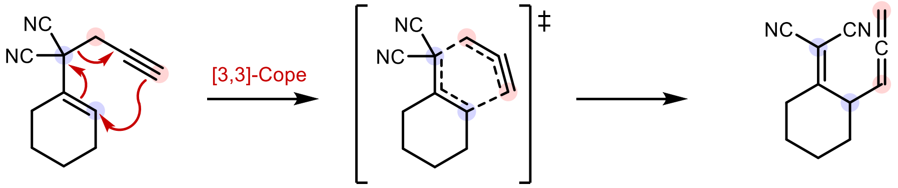
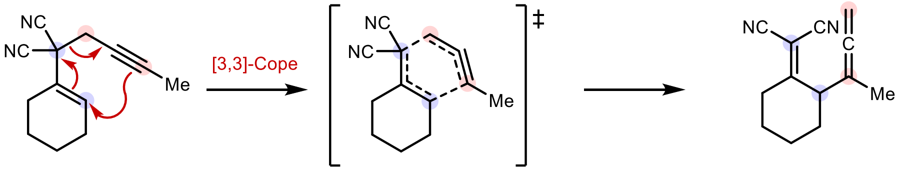
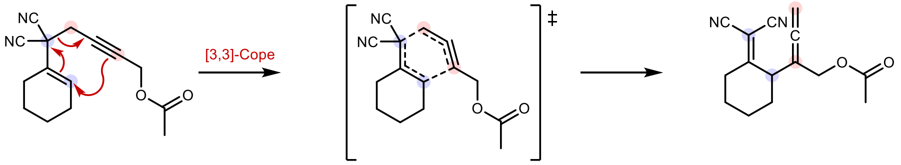
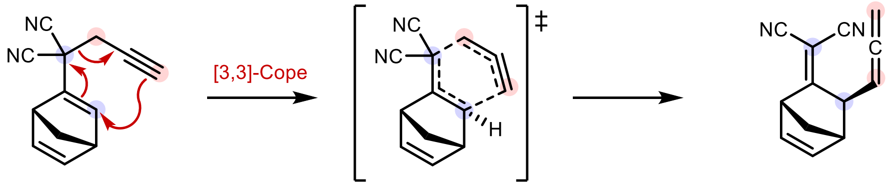
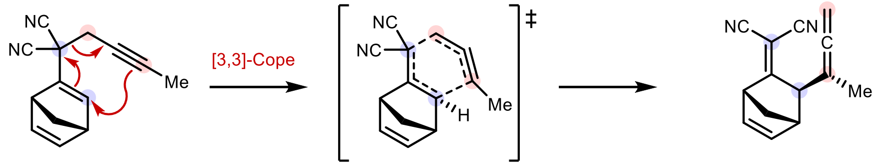
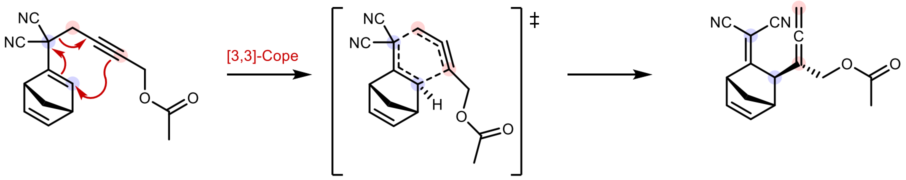
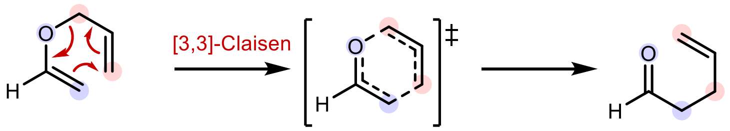
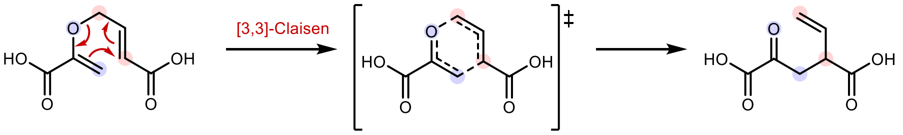
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
79		1.94	42.25	-40.31
80		28.38	41.25	-12.87
81		11.14	43.06	-31.92
82		10.61	37.95	-27.34
83		17.65	44.98	-27.33
84		26.68	50.51	-23.83
85		30.66	51.43	-20.78
86		27.74	52.70	-24.96
87		30.73	53.65	-22.91

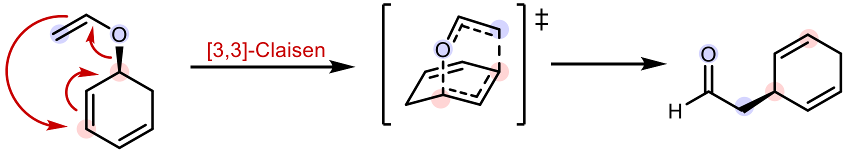
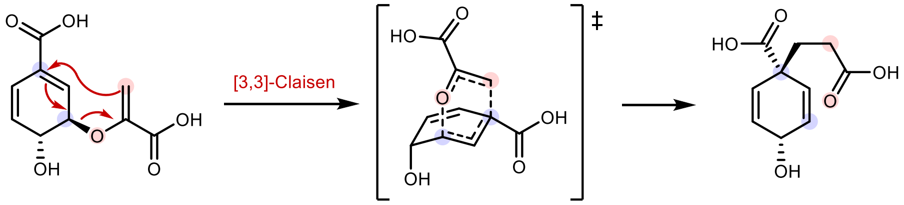
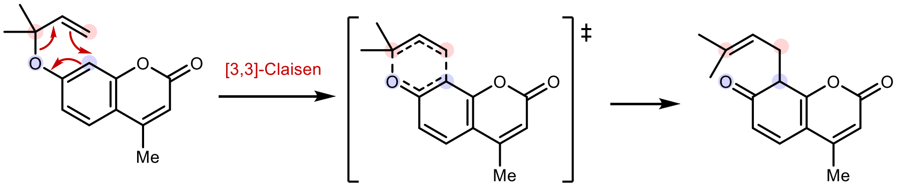
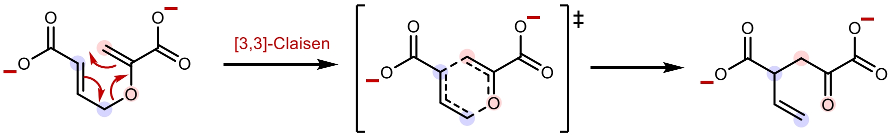
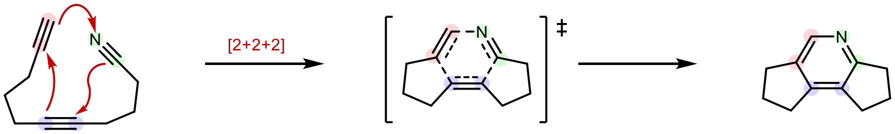
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
88		24.91	51.36	-26.45
89		28.62	52.88	-24.26
90		27.43	52.17	-24.74
91		31.00	54.55	-23.55
92		23.66	53.48	-29.82
93		28.69	54.95	-26.26
94		25.60	54.46	-28.86
95		28.02	55.88	-27.86
96		11.39	15.19	-3.80

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
97		43.52	21.76	21.76
98		33.08	45.42	-12.34
99		44.28	54.35	-10.07
100	 H N H Electrocyclic N ‡ H H N H H	36.07	19.79	16.28
101	 H N H Electrocyclic N ‡ H H N H H	41.80	45.81	-4.01
102		38.73	36.55	2.18
103	 H H N H H Electrocyclic N ‡ H N H	36.24	36.68	-0.44
104	 O S O H H H H Electrocyclic O S O ‡ H O S O H	13.10	40.49	-27.39
105	 H H O S O H H Electrocyclic O S O ‡ H O S O H	48.25	42.24	6.01
106	 NC CN Me Me [4+6] NC CN Me Me ‡ NC CN Me Me	11.53	29.79	-18.26

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
107	 NC CN Me Me [6+4] CN CN H ‡ CN CN H	11.53	36.38	-24.85
108	 NC CN Me Me [8+2] NC CN Me Me ‡ NC CN Me Me	11.53	35.01	-23.48
109	 C F F F F [3,3] C F F F F ‡ C F F F F	43.45	50.58	-7.13
110	 F F C F F [3,3] F F C F F ‡ F F C F F	43.45	57.37	-13.92
111	 C H H H H [3,3] C H H H H ‡ C H H H H	36.96	51.17	-14.22
112	 MeO2C CN Me Me [4+6] MeO2C CN Me Me ‡ NC CO2Me Me Me	12.84	31.12	-18.28
113	 MeO2C CN Me Me [6+4] CO2Me CN H ‡ CO2Me CN H	12.84	36.75	-23.91
114	 MeO2C CN Me Me [8+2] MeO2C CN Me Me ‡ MeO2C CN Me Me	12.84	37.94	-25.10
115	 NC CO2Me Me Me [4+6] NC CO2Me Me Me ‡ MeO2C CN Me Me	11.83	30.11	-18.28

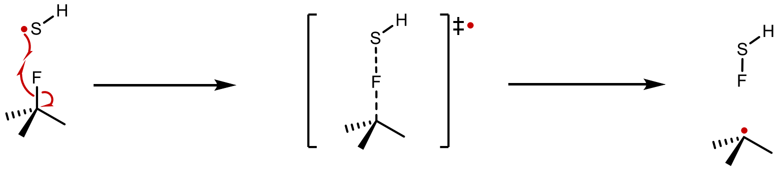
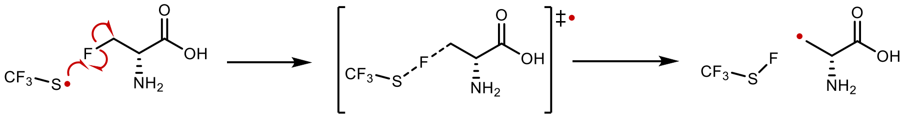
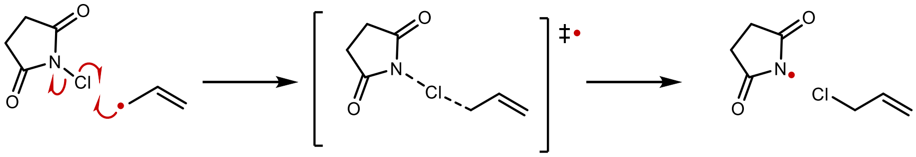
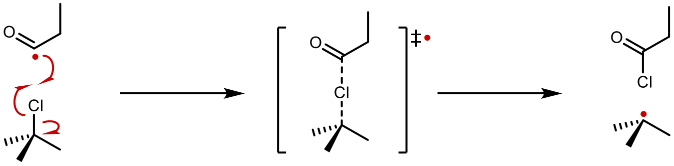
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
116		11.83	35.81	-23.98
117		11.83	40.21	-28.38
118		22.52	41.61	-19.09
119		18.91	47.58	-28.67
120		16.79	43.12	-26.33
121		16.79	54.33	-37.54
122		19.29	37.46	-18.17
123		19.29	41.95	-22.66
124		21.68	46.02	-24.34
125		21.68	47.32	-25.64

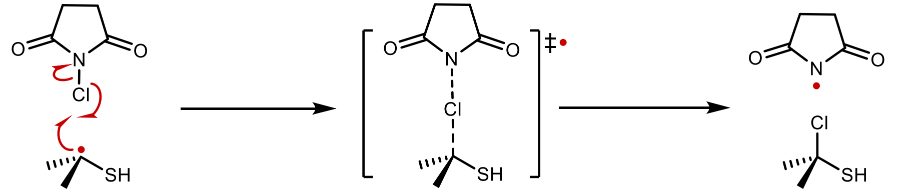
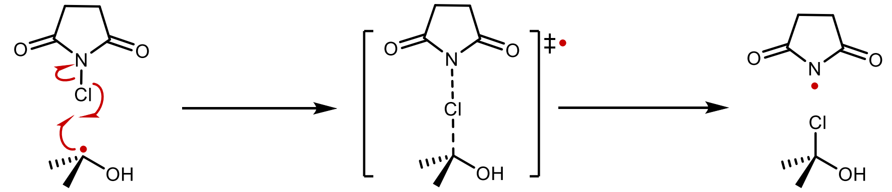
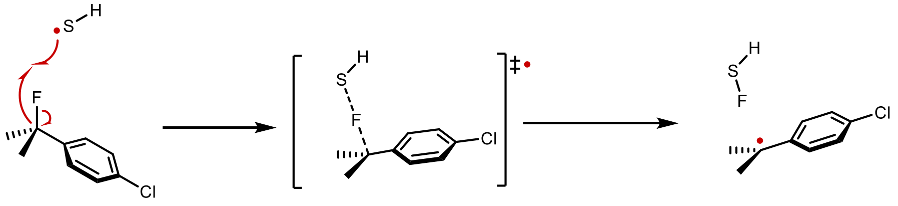
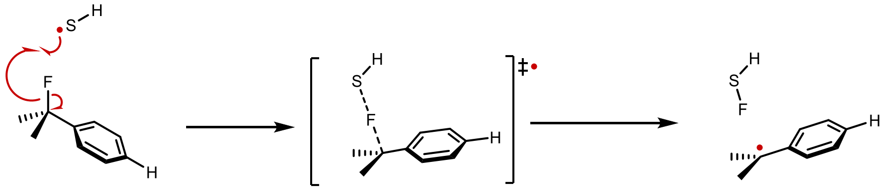
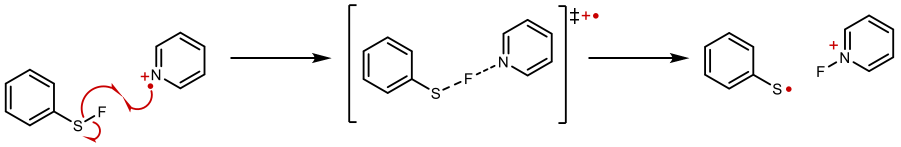
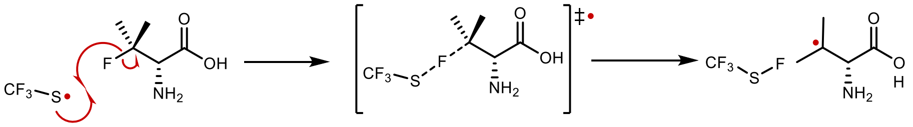
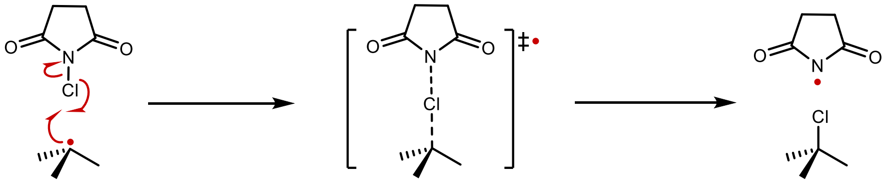
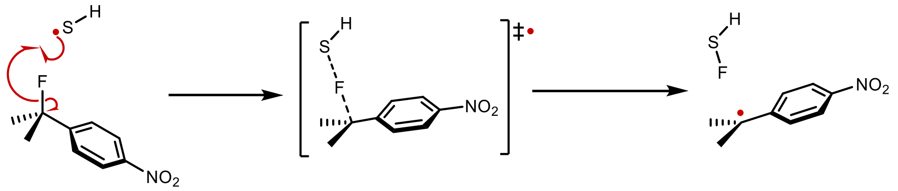
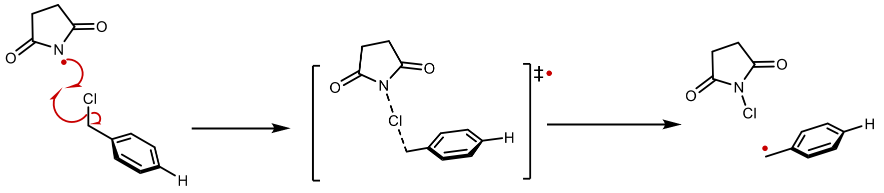
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
126		20.25	26.19	-5.95
127		20.95	32.16	-11.21
128		32.74	41.45	-8.71
129		34.42	40.15	-5.73
130		35.40	43.32	-7.92
131		32.39	43.38	-10.99
132		32.92	41.10	-8.18
133		33.84	44.83	-10.99
134		32.08	49.61	-17.53
135		28.35	44.42	-16.07

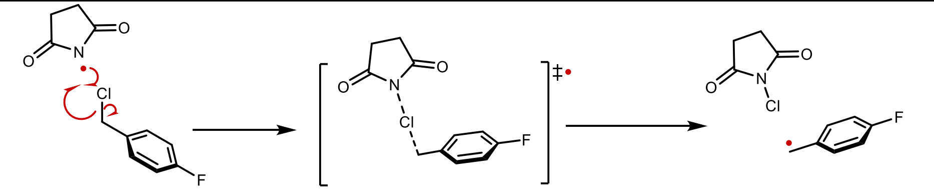
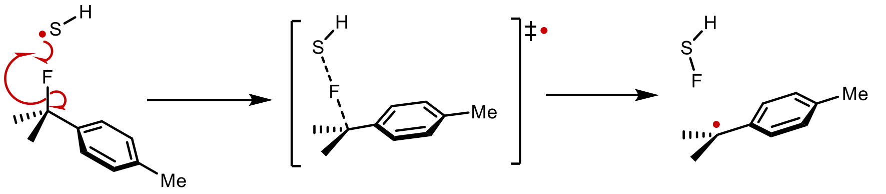
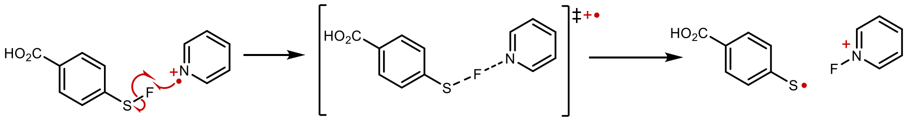
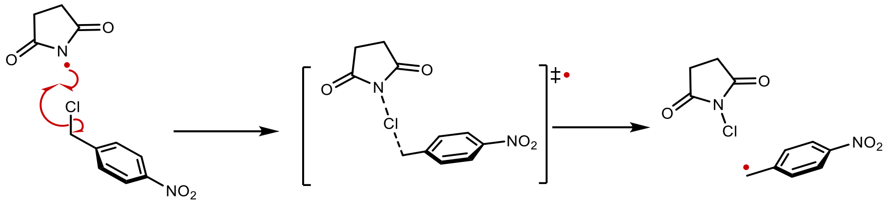
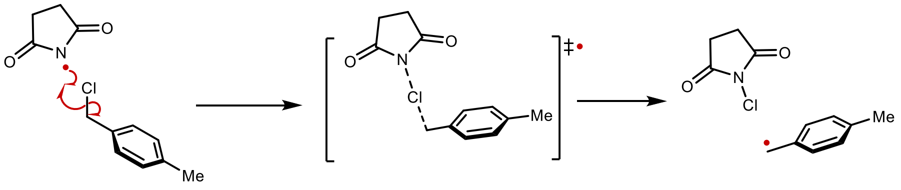
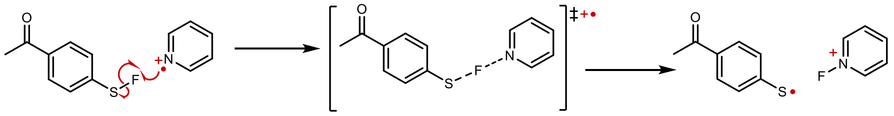
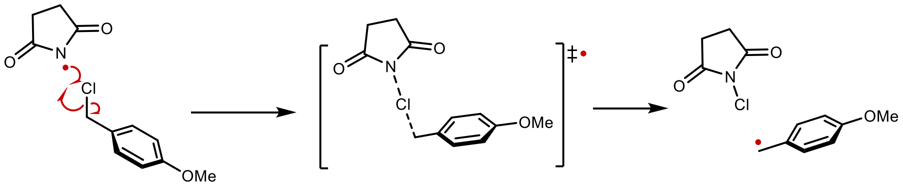
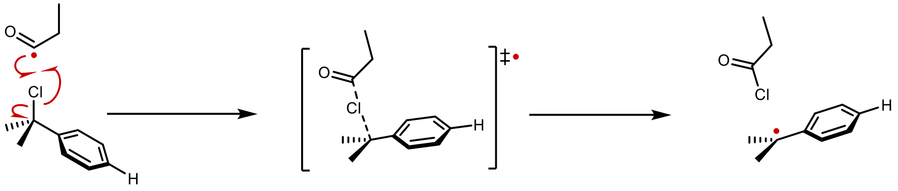
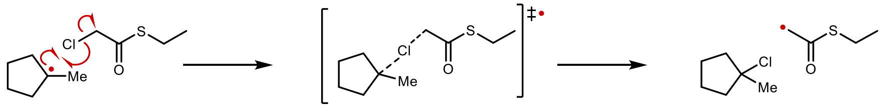
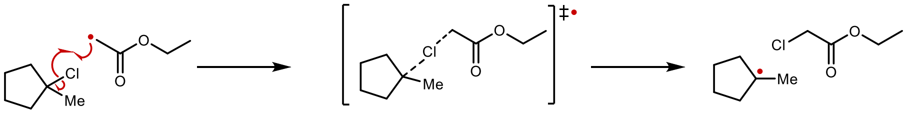
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
136		29.79	45.83	-16.04
137	 O OH O OH OH O [3,3]-Claisen HO O O OH OH O ‡ HO O OH O OH	25.98	41.35	-15.37
138	 O O O Me [3,3]-Claisen O O O Me ‡ O O O Me	31.94	28.09	3.85
139	 O O O O O [3,3]-Claisen O O O O O ‡ O O O O O	34.37	50.96	-16.59
140		54.84	144.39	-89.55

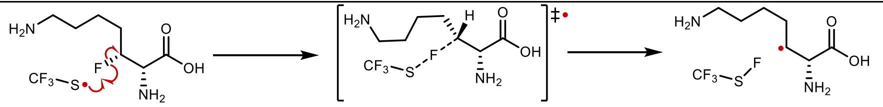
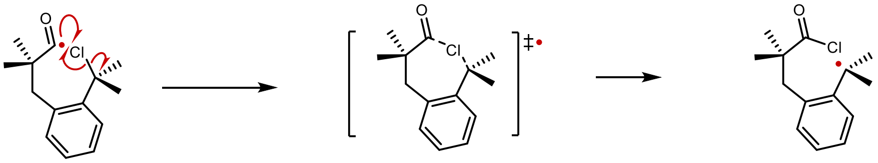
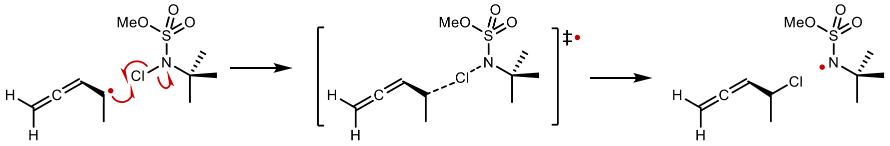
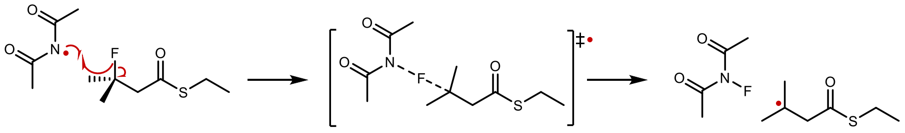
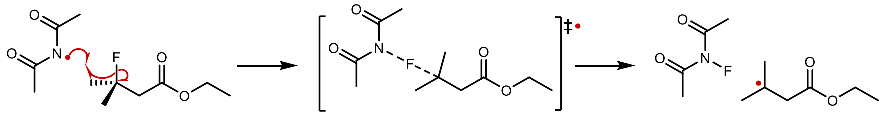
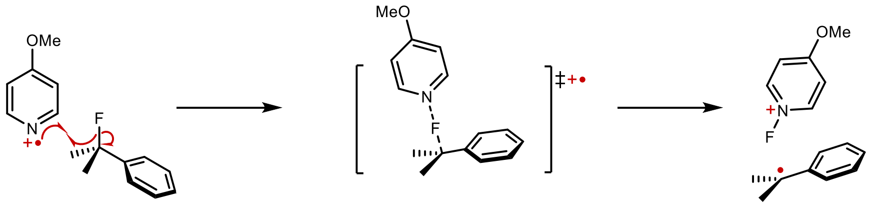
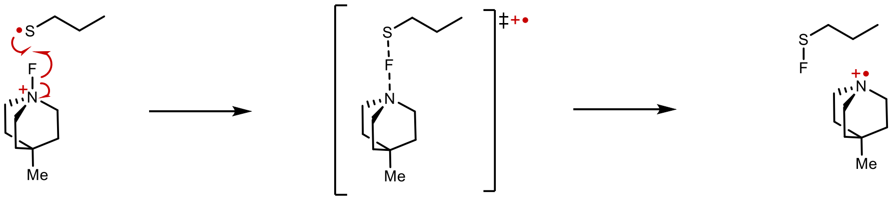
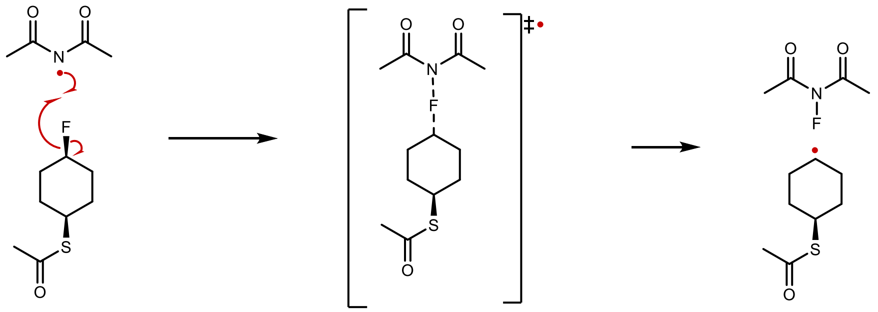
[a] units are in kcal/mol

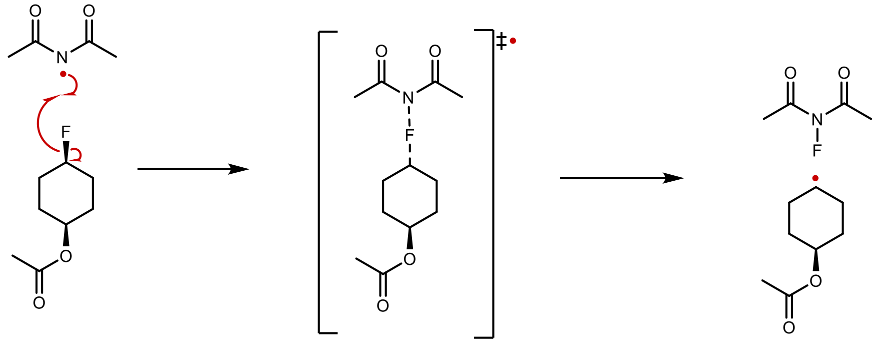
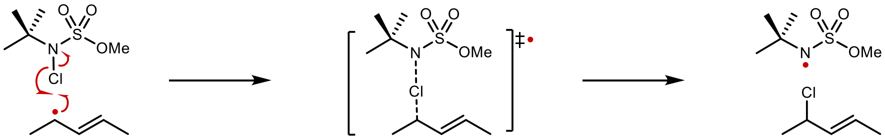
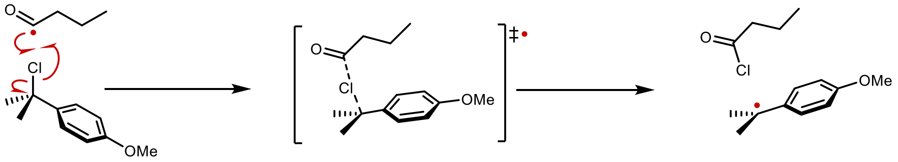
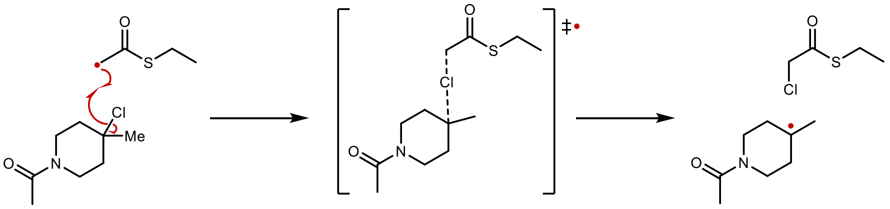
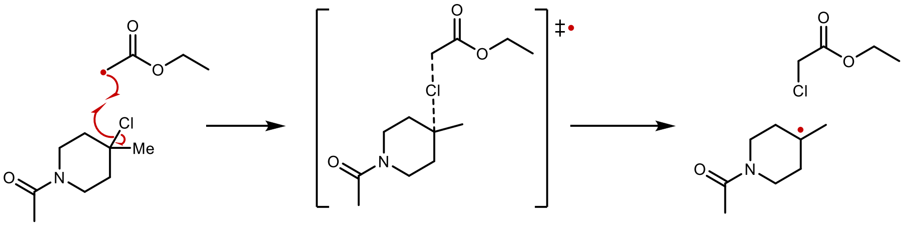
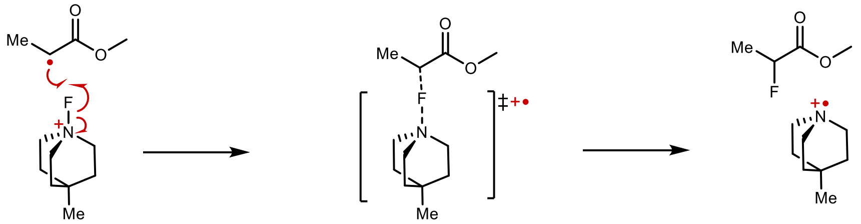
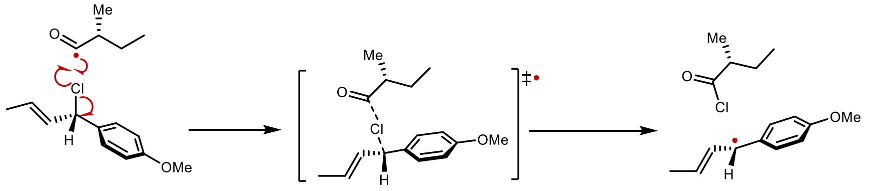
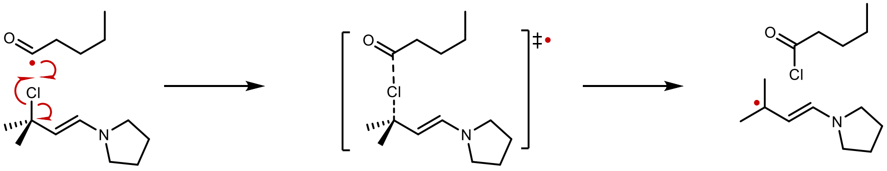
III. Halogen atom transfer

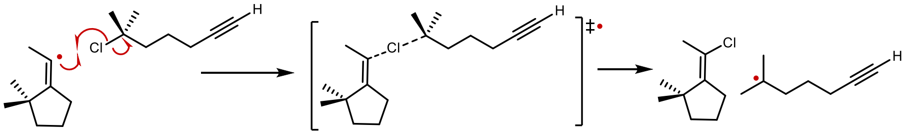
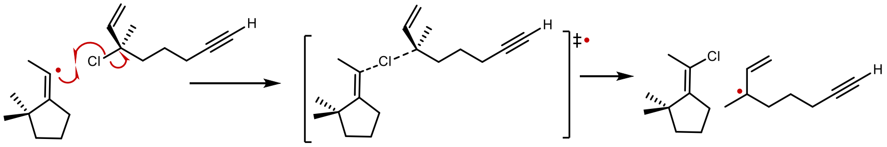
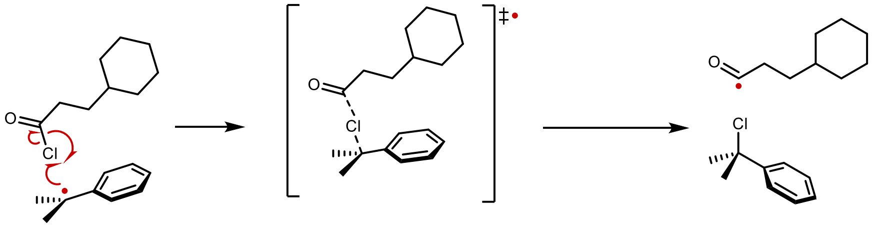
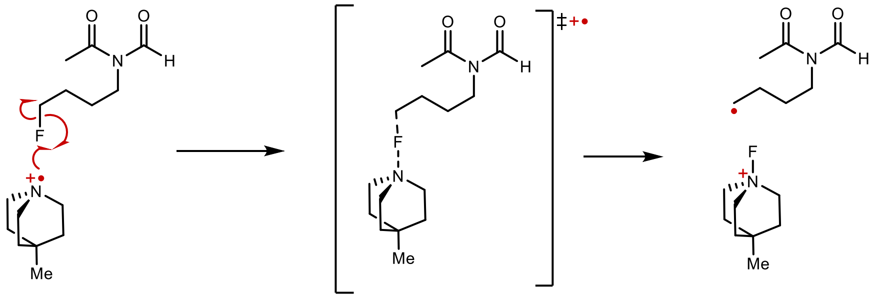
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
1		47.62	11.46	36.16
2	 CF₃ S F O OH NH₂ CF₃ S F O OH NH₂ ‡• CF₃ S F O OH NH₂	45.08	13.69	31.39
3	 O N Cl O O N Cl O ‡• O N O Cl	15.39	4.08	11.31
4		15.91	13.98	1.93

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
5		4.23	4.65	-0.42
6		-1.35	1.80	-3.16
7		42.38	17.76	24.62
8		42.36	17.42	24.94
9		-67.71	-62.16	-5.55
10		41.34	7.92	33.42
11		0.50	5.21	-4.70
12		43.97	20.13	23.84
13		3.23	11.60	-8.37

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
14		3.48	11.59	-8.11
15		41.94	16.76	25.18
16		77.42	81.94	-4.52
17		9.22	14.73	-5.52
18		2.95	11.10	-8.15
19		-0.68	3.87	-4.54
20		2.41	10.55	-8.14
21		12.96	23.34	-10.38
22		10.47	18.30	-7.83
23		18.66	10.80	7.85

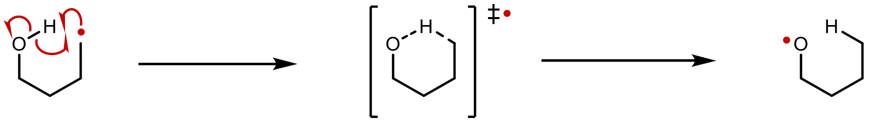
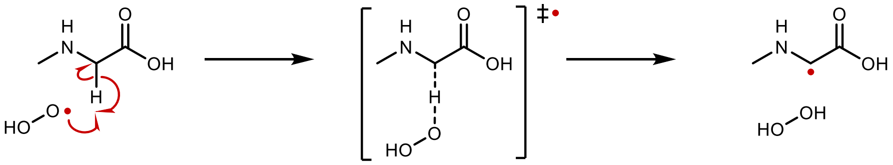
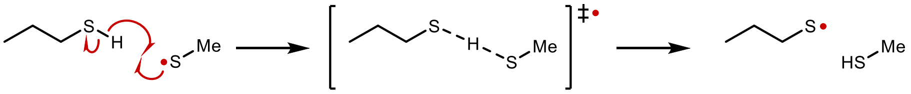
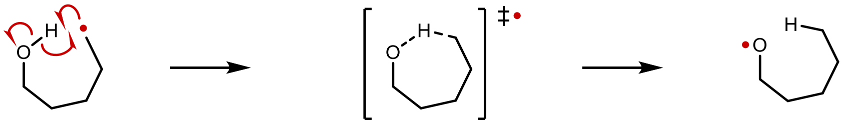
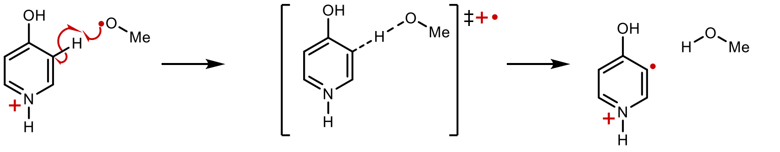
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
24		41.74	8.44	33.31
25		26.26	34.43	-8.17
26		5.84	26.05	-20.21
27		53.90	10.31	43.59
28		56.10	11.62	44.48
29		26.77	1.36	25.42
30		14.71	34.21	-19.50
31		51.57	12.40	39.17

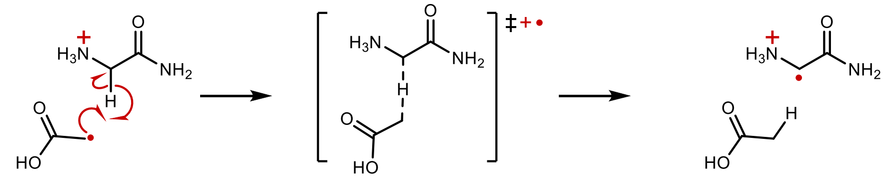
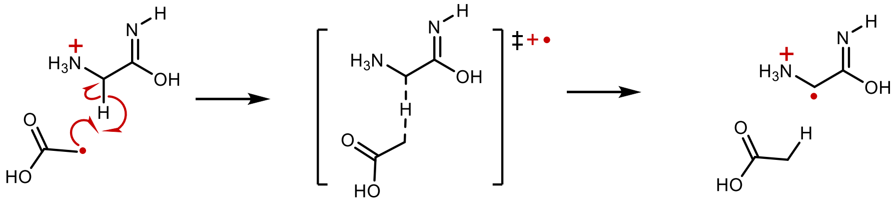
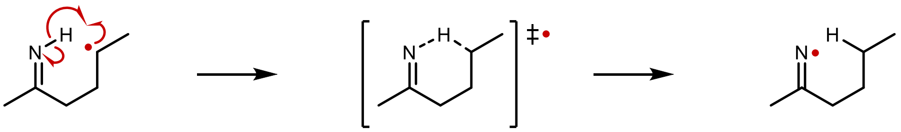
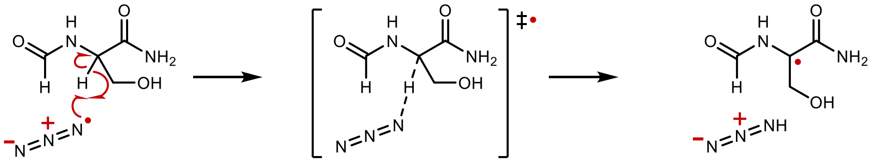
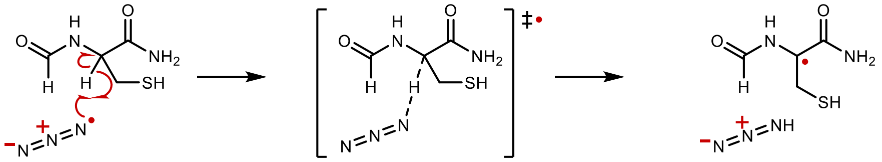
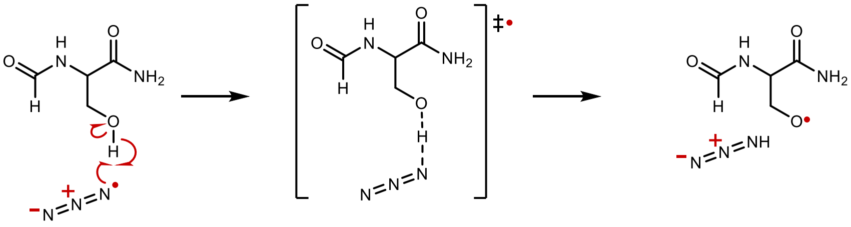
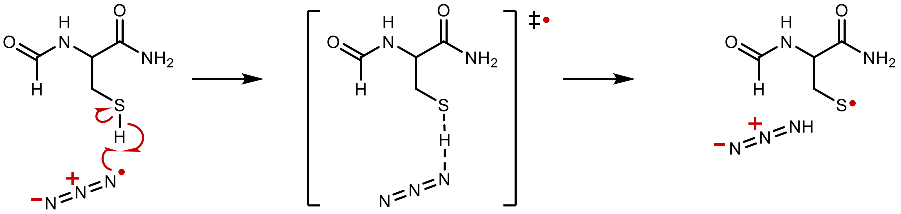
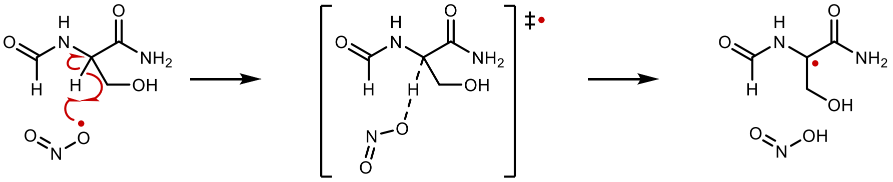
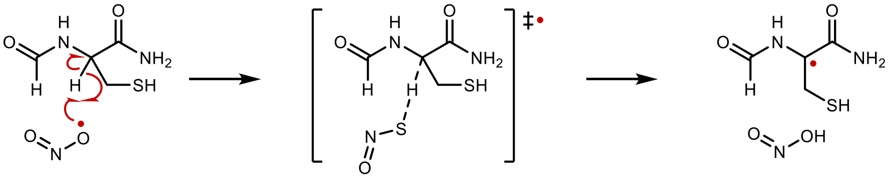
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
32		52.03	12.75	39.29
33		6.22	25.32	-19.10
34		12.79	22.60	-9.81
35		16.87	8.88	7.99
36		16.65	8.64	8.02
37		8.00	47.09	-39.09
38		10.67	31.39	-20.71
39		11.00	24.57	-13.57

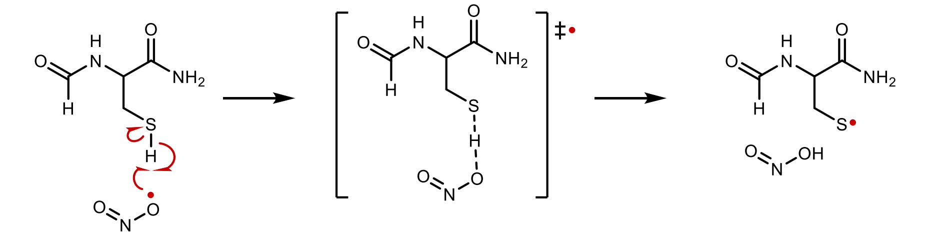
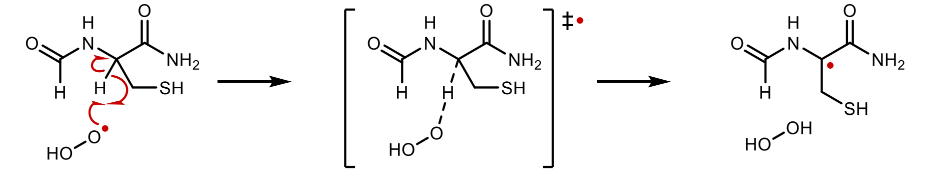
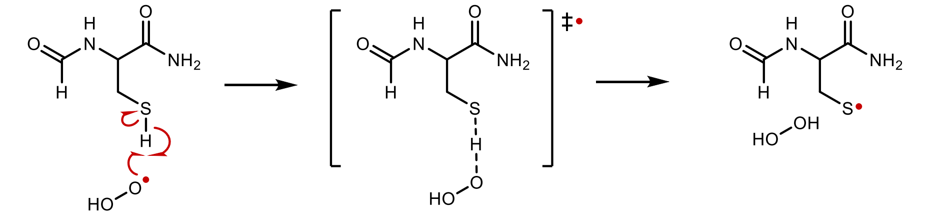
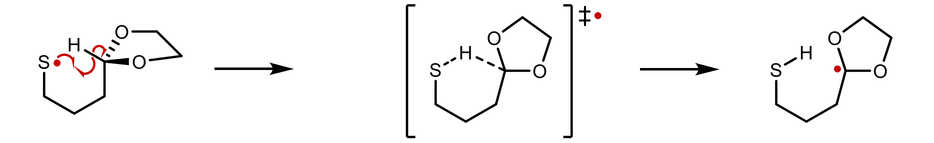
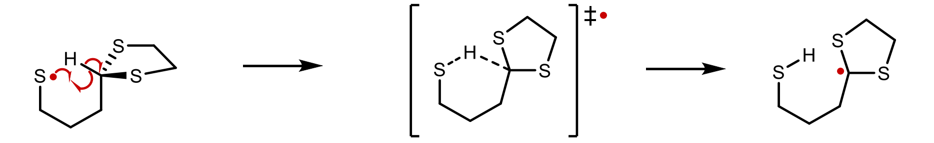
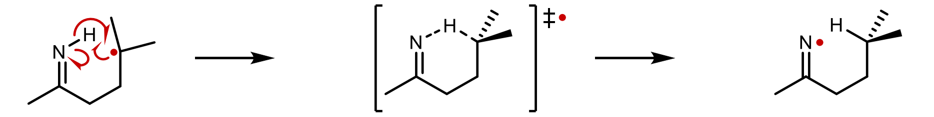
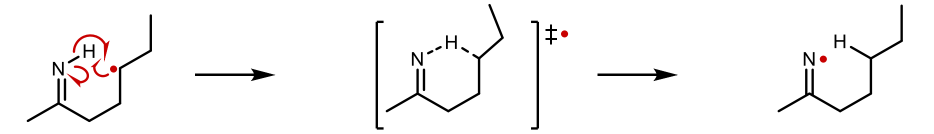
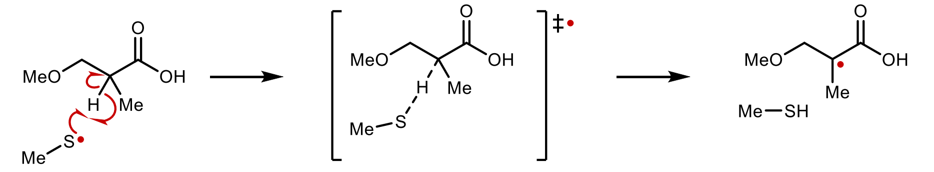
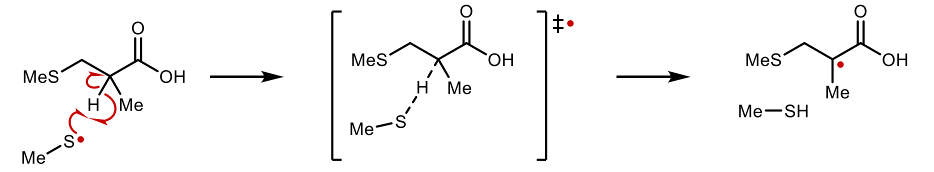
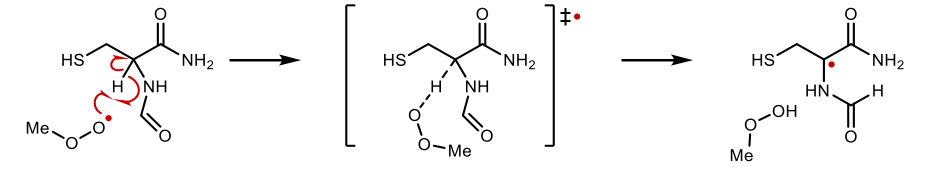
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
40		10.04	20.95	-10.91
41		7.96	35.40	-27.43
42		21.33	10.29	11.03
43		46.53	-3.19	49.72

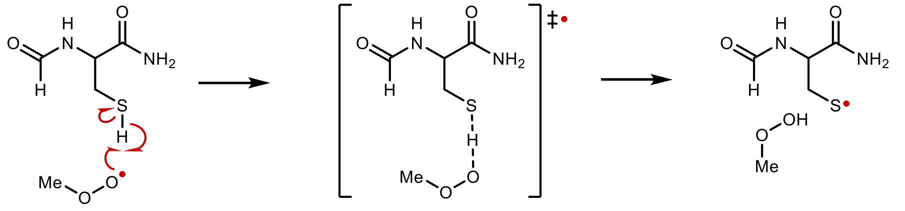
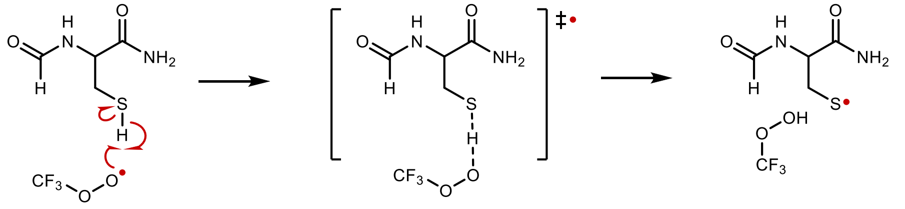
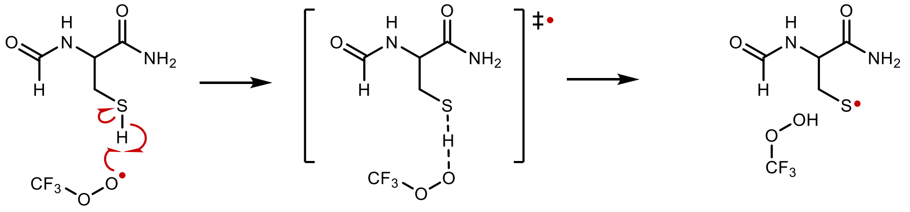
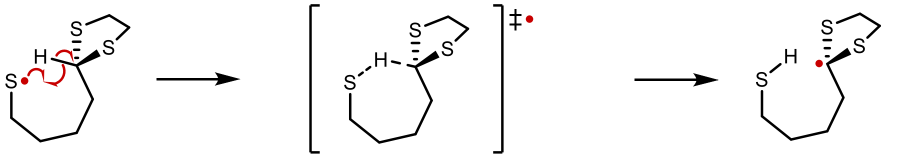
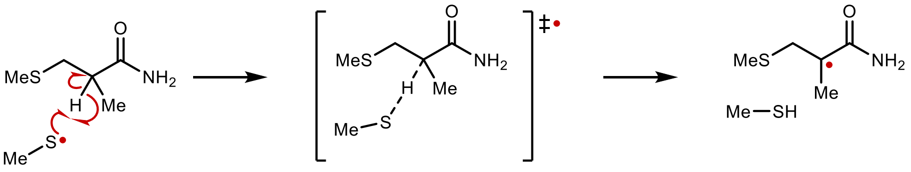
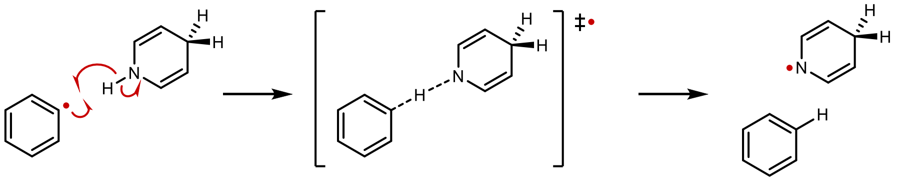
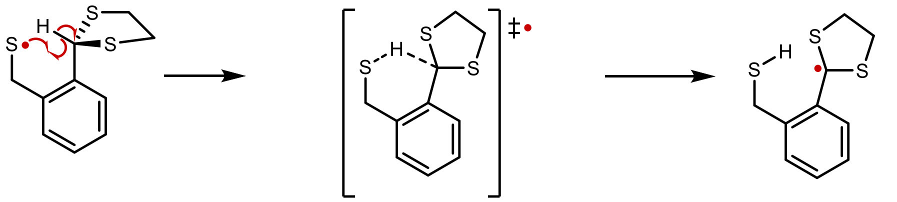
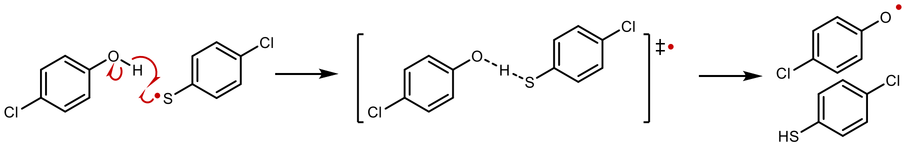
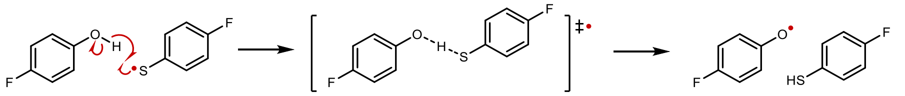
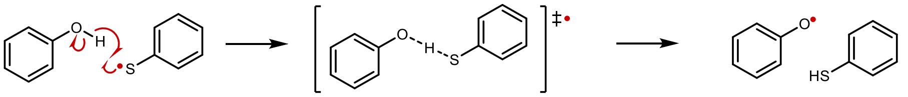
[a] units are in kcal/mol.

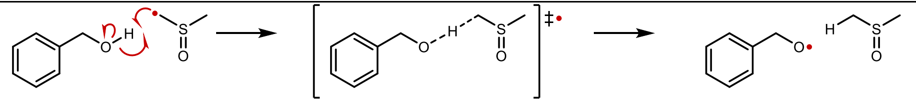
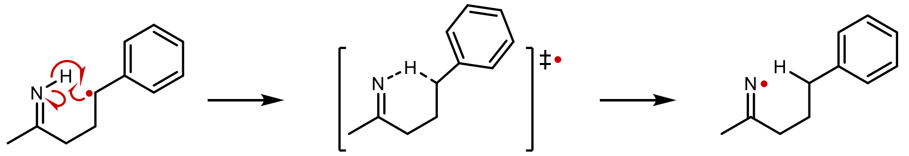
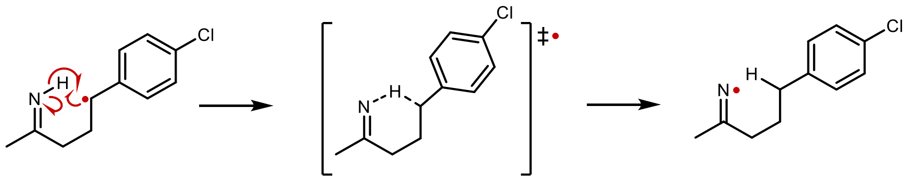
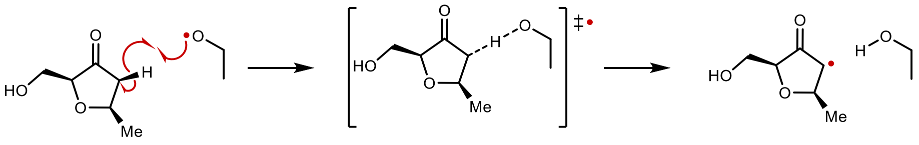
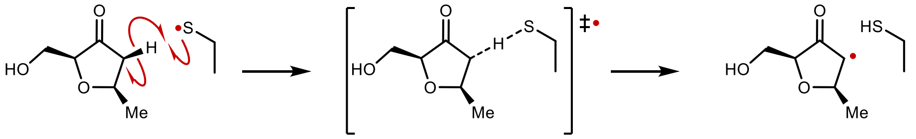
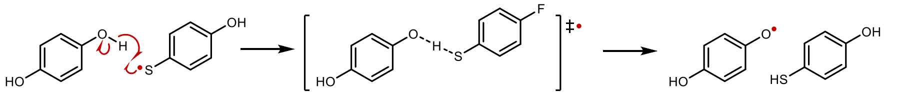
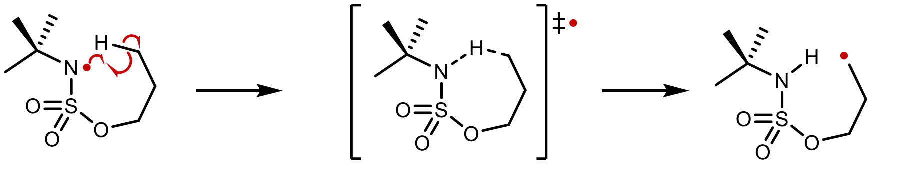
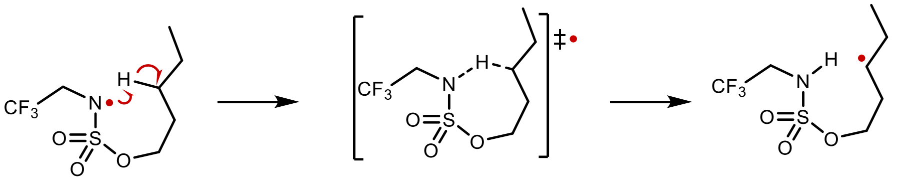
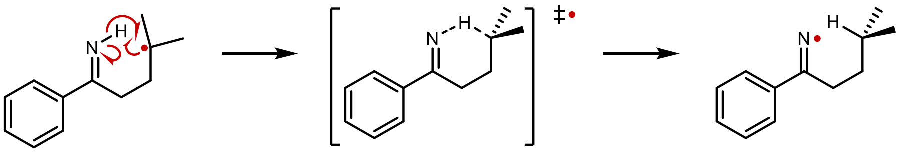
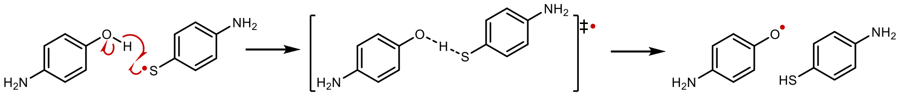
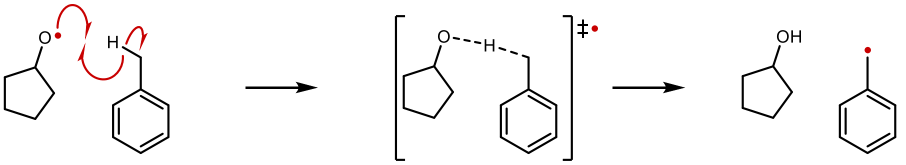
IV. Hydrogen atom transfer

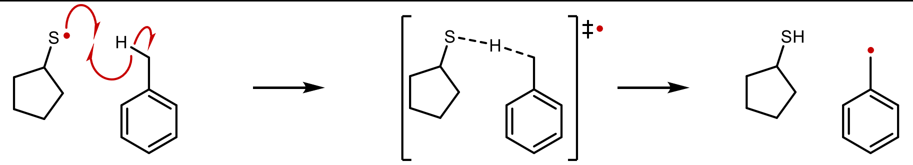
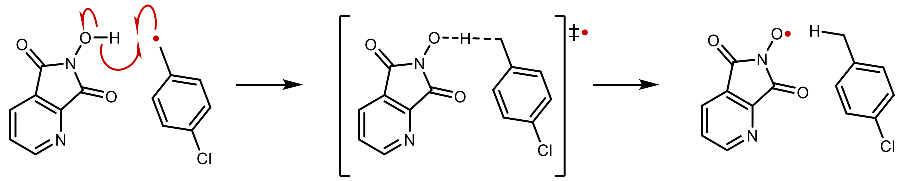
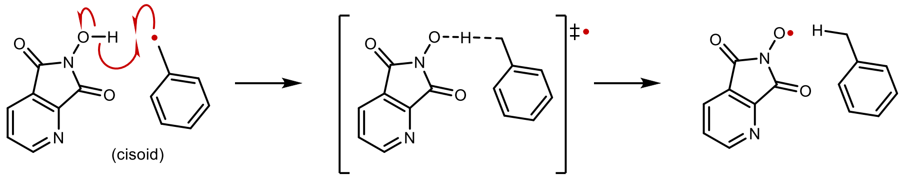
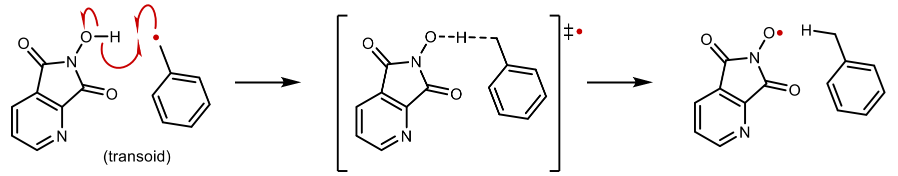
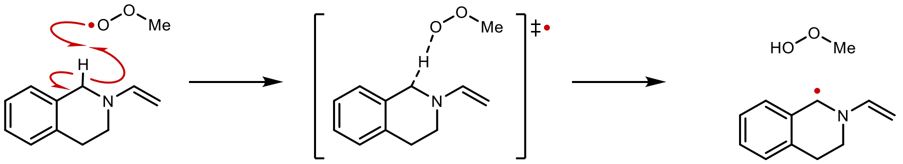
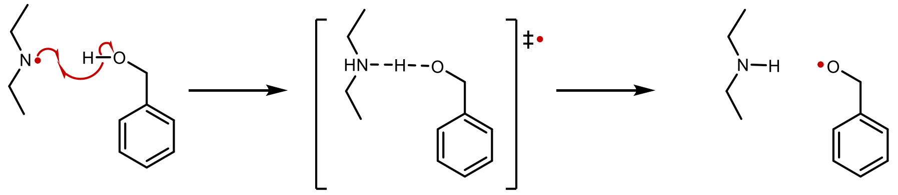
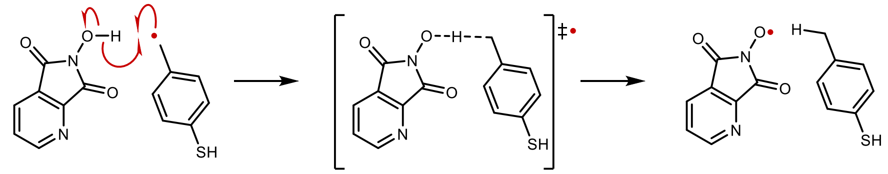
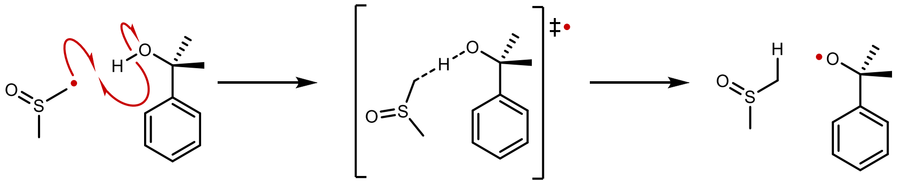
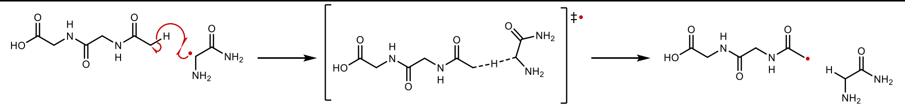
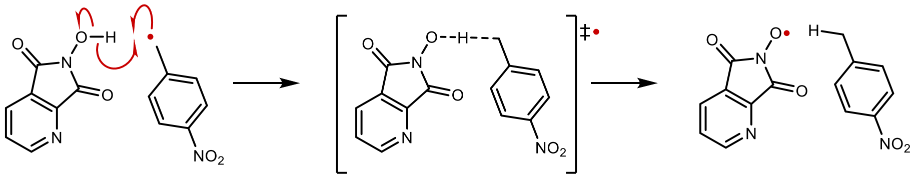
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
1		16.42	12.36	4.06
2		11.89	19.47	-7.58
3		4.80	4.58	0.22
4		16.44	12.67	3.77
5		13.93	0.17	13.76

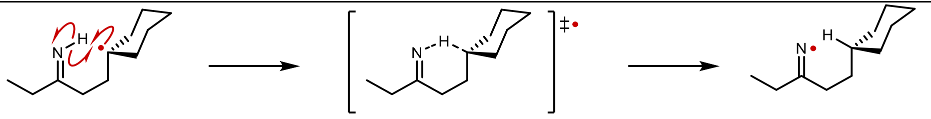
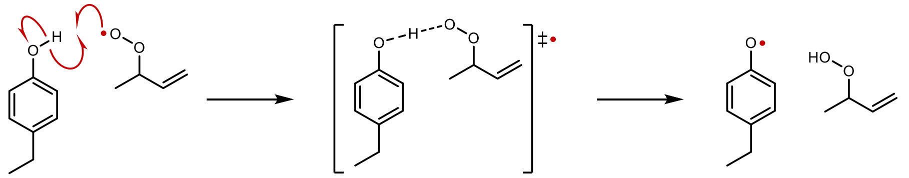
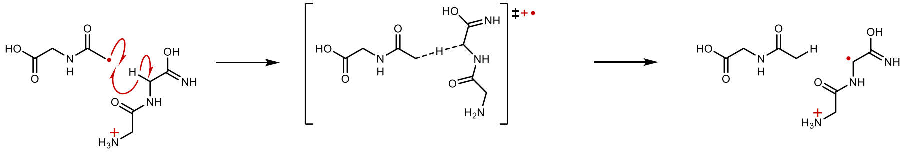
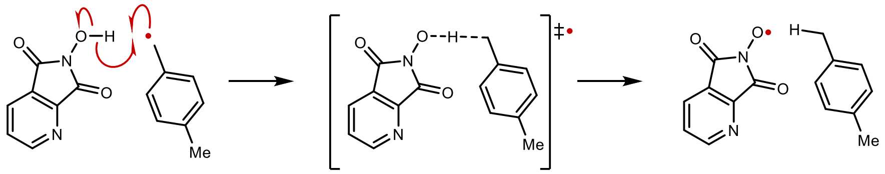
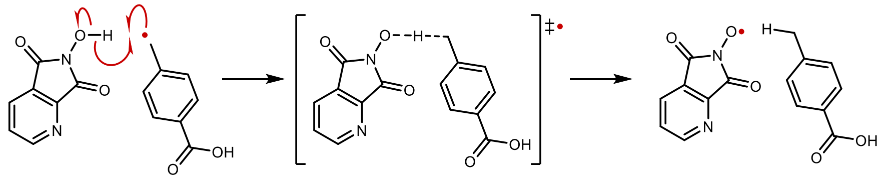
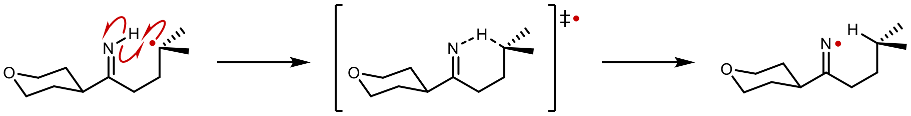
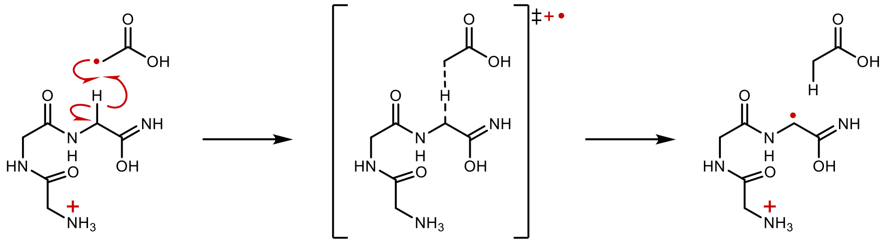
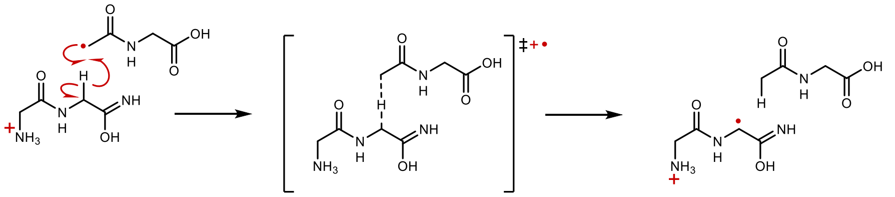
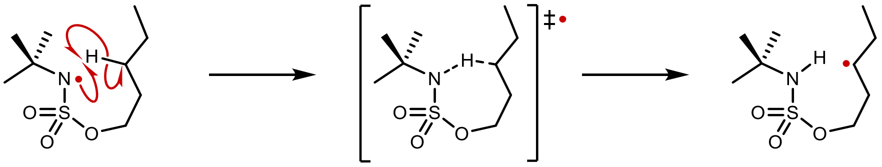
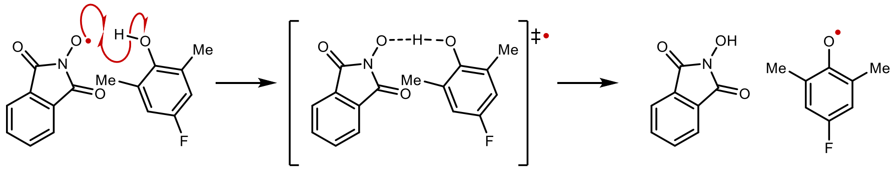
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
6		6.44	2.44	4.01
7		4.00	4.61	-0.61
8		12.36	21.11	-8.75
9		17.45	17.15	0.30
10		16.29	17.84	-1.56
11		34.10	12.51	21.59
12		12.97	15.74	-2.77
13		30.51	19.78	10.73
14		27.87	18.89	8.98

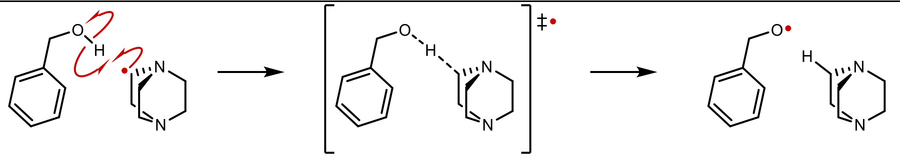
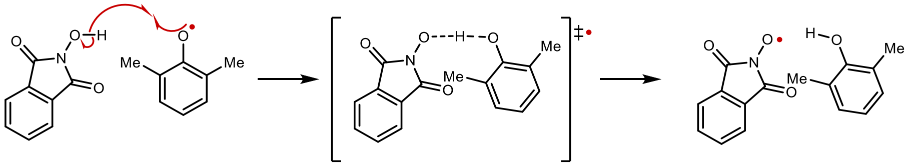
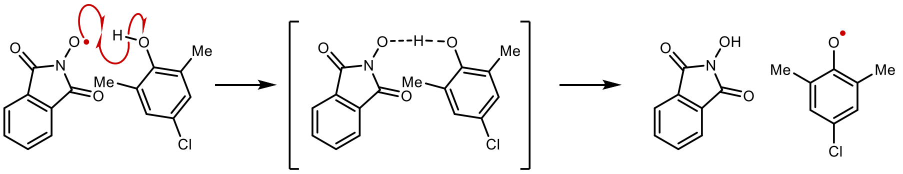
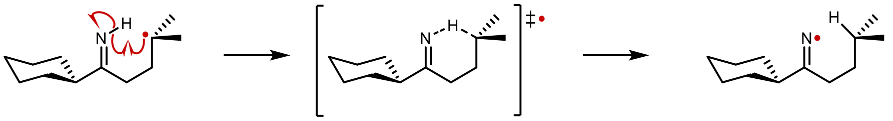
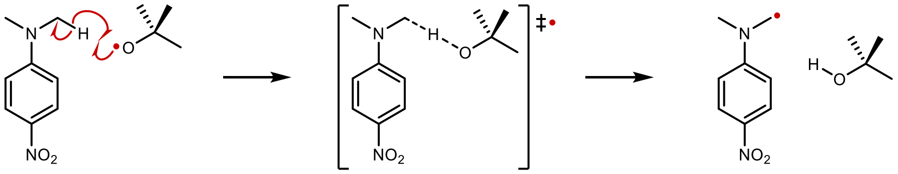
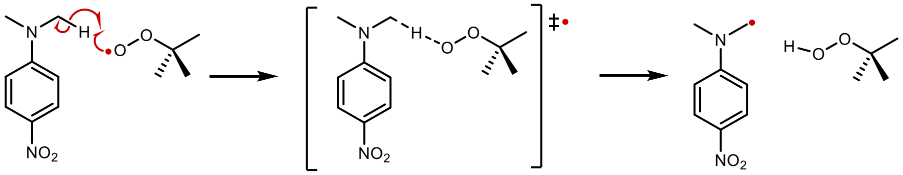
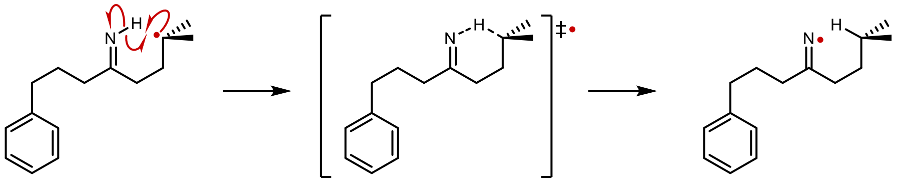
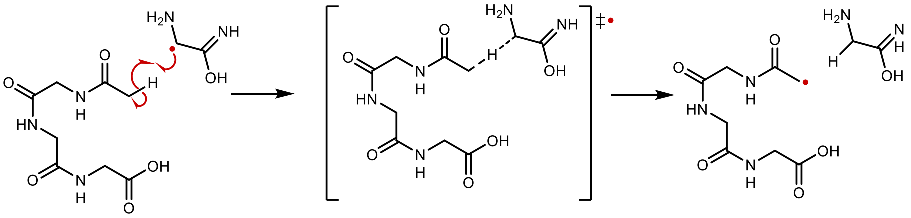
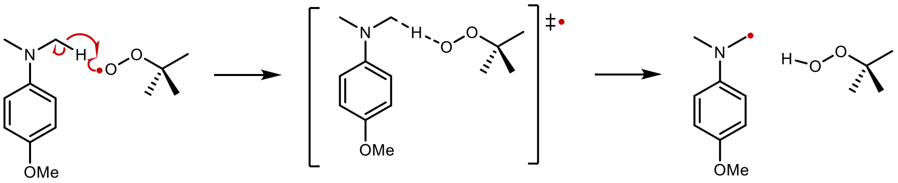
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
15		23.98	16.31	7.67
16		14.52	13.49	1.02
17		9.08	9.37	-0.29
18		12.84	3.23	9.61
19		10.35	9.72	0.63
20		12.19	19.25	-7.06
21		11.81	20.88	-9.06
22		13.47	6.04	7.43
23		11.29	6.86	4.43
24		18.51	15.80	2.71

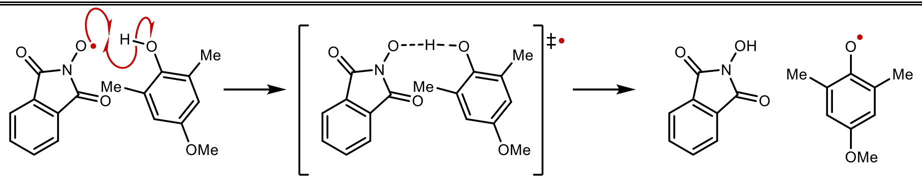
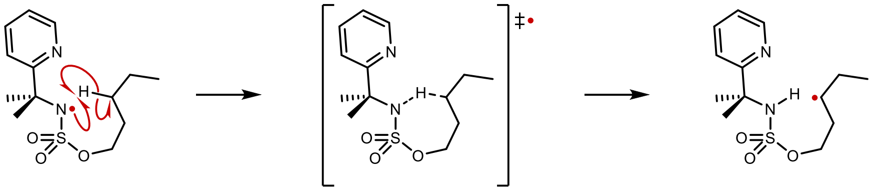
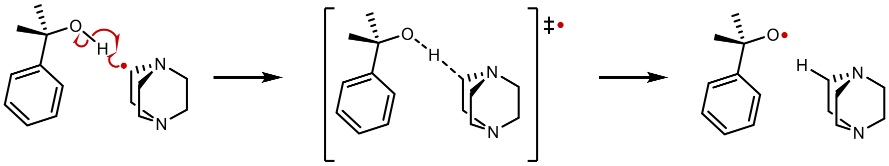
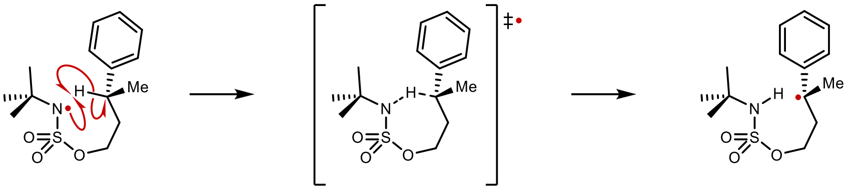
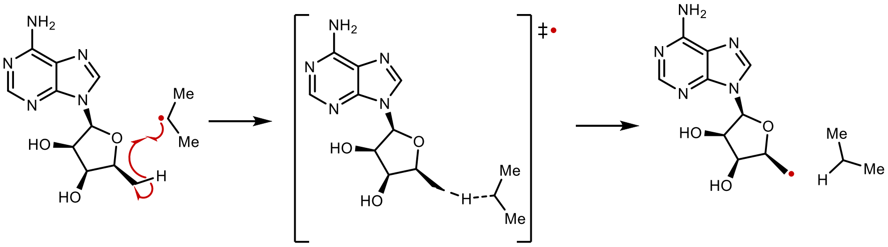
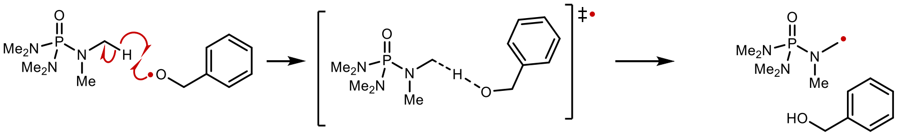
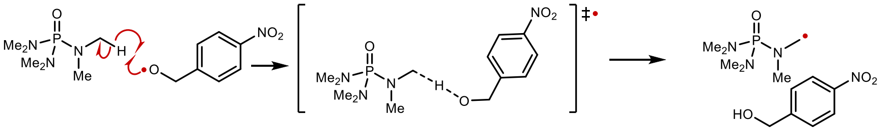
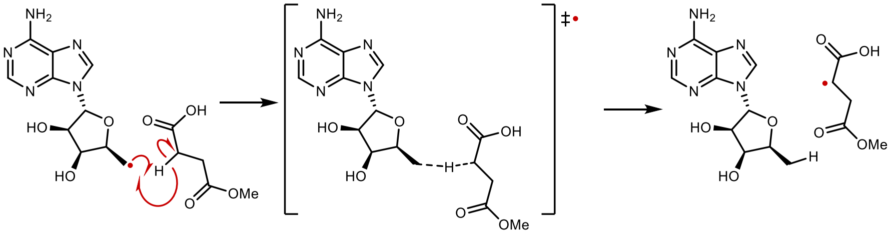
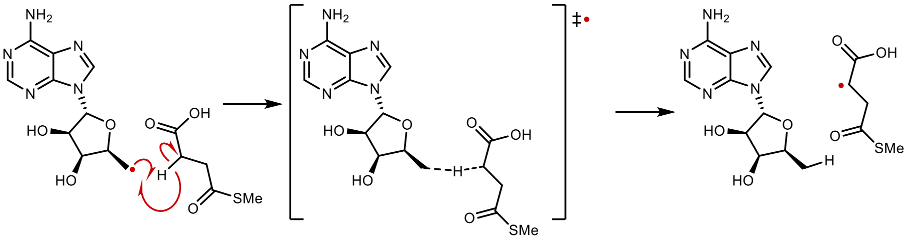
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
25		14.12	12.73	1.39
26		14.83	20.51	-5.68
27		13.20	18.89	-5.69
28		10.10	8.71	1.38
29		11.34	5.16	6.19
30		3.48	35.56	-32.09
31		10.13	11.96	-1.83
32		29.28	18.90	10.38
33		30.75	20.43	10.31
34		31.27	20.17	11.10

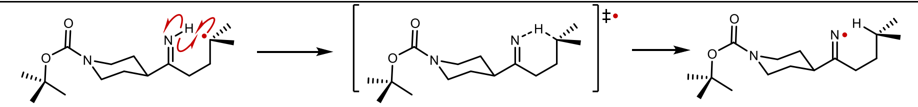
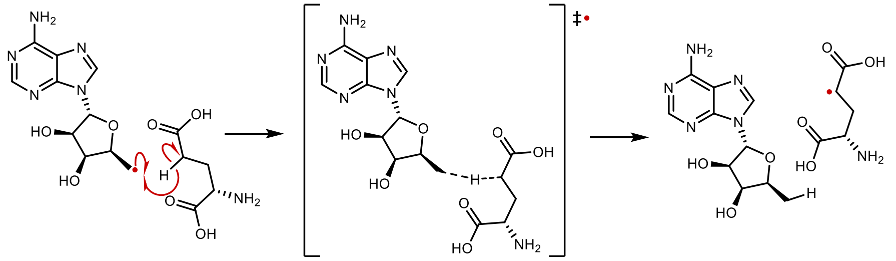
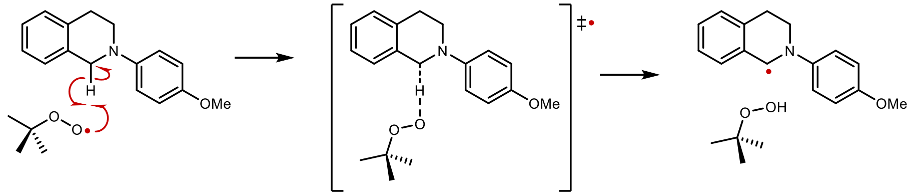
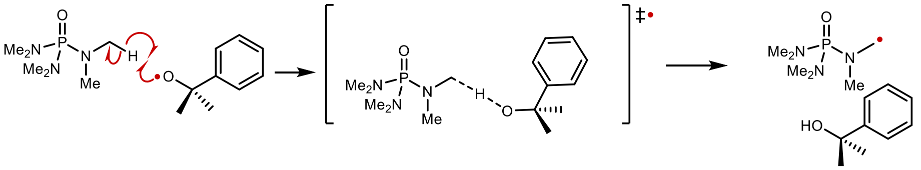
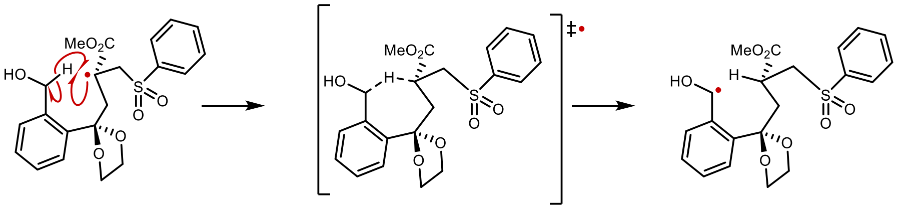
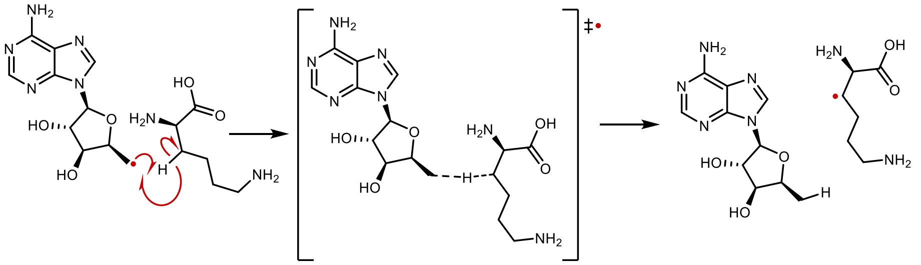
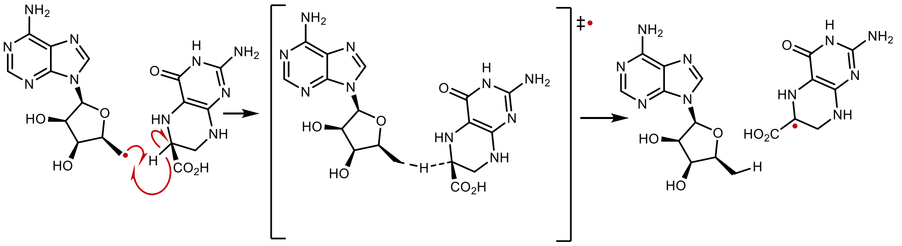
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
35		12.23	7.14	5.09
36		21.30	19.97	1.33
37		21.57	19.91	1.66
38		6.68	17.92	-11.24
39		13.24	3.28	9.96
40		27.97	19.50	8.47
41		18.43	18.66	-0.24
42		12.15	16.10	-3.95
43		13.68	18.76	-5.08
44		22.68	15.83	6.85
45		7.33	22.38	-15.05

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
46		11.46	6.32	5.14
47		5.60	11.33	-5.72
48		5.93	11.61	-5.68
49		12.58	18.27	-5.69
50		12.73	18.65	-5.92
51		11.99	-0.34	12.34
52		4.82	9.92	-5.10
53		12.35	7.34	5.02
54		19.94	2.71	17.22
55		6.94	12.94	-6.00

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
56		12.91	18.25	-5.35
57		14.38	12.67	1.71
58		12.42	2.71	9.71
59		5.02	10.66	-5.64
60		6.68	12.22	-5.55
61		11.91	19.00	-7.09
62		25.06	34.50	-9.44
63		12.43	20.69	-8.26
64		14.37	17.61	-3.24
65		3.46	3.34	0.12

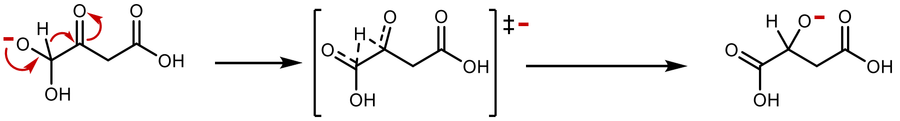
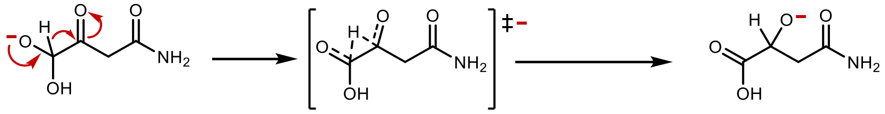
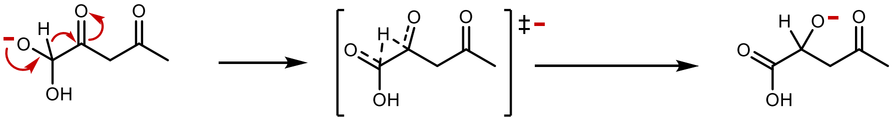
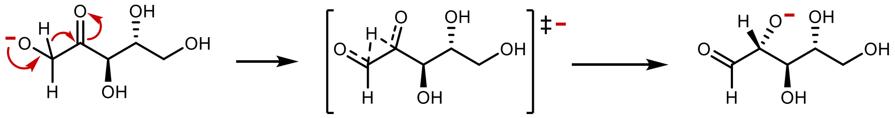
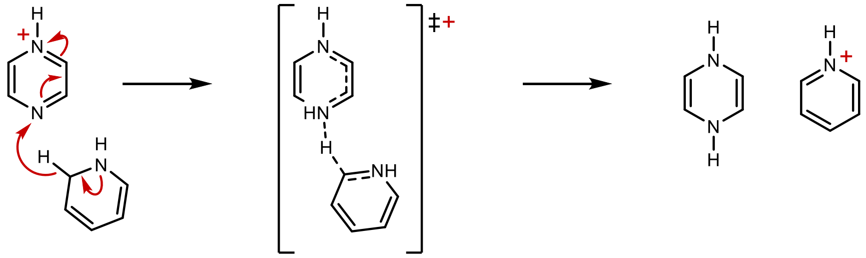
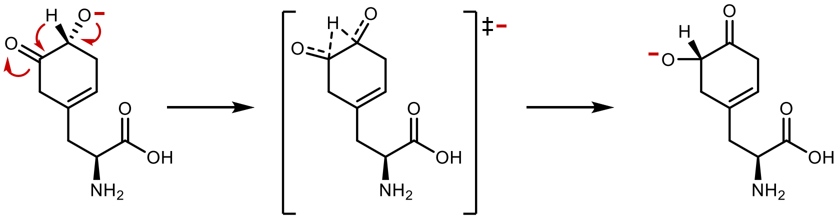
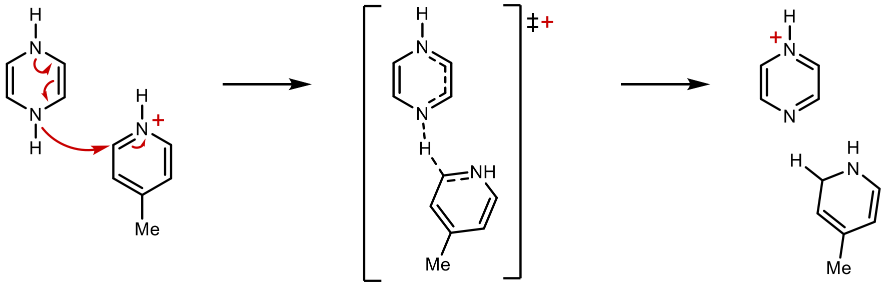
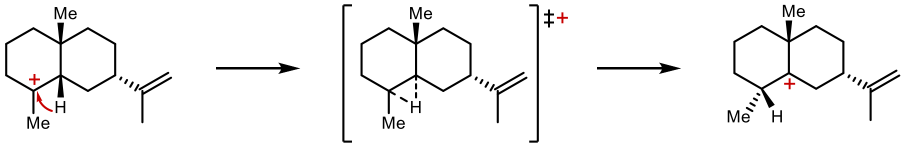
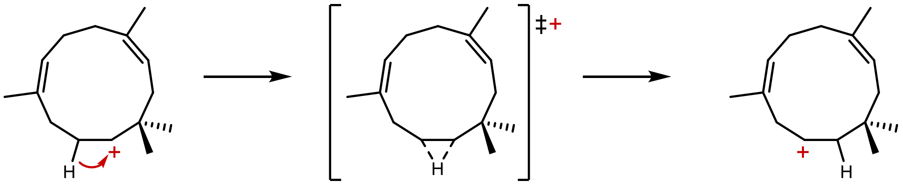
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
66		12.91	5.36	7.55
67		3.55	5.23	-1.69
68		3.62	2.75	0.87
69		11.97	19.19	-7.22
70		5.16	19.31	-14.15
71		16.88	8.08	8.80
72		11.57	19.48	-7.91
73		30.00	12.73	17.28
74		14.37	7.11	7.26

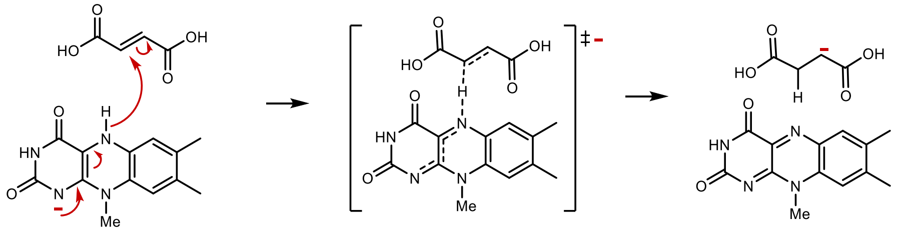
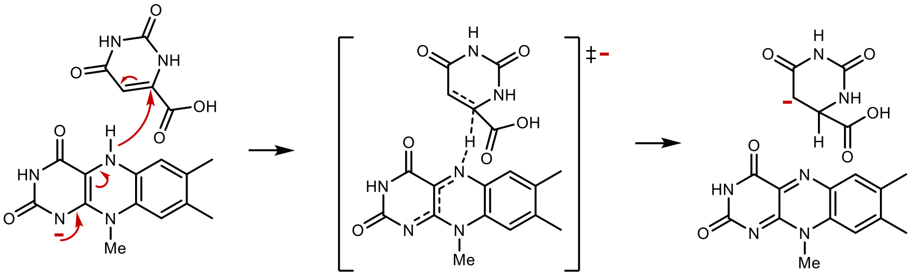
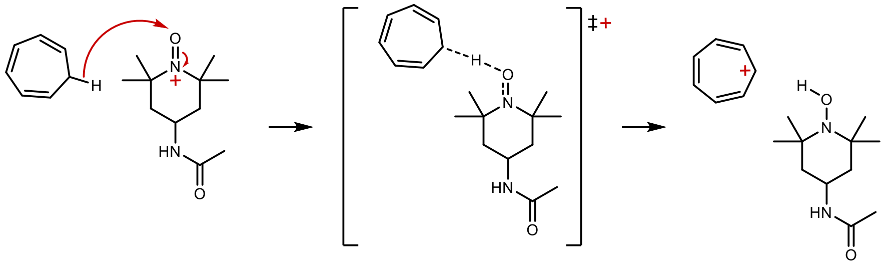
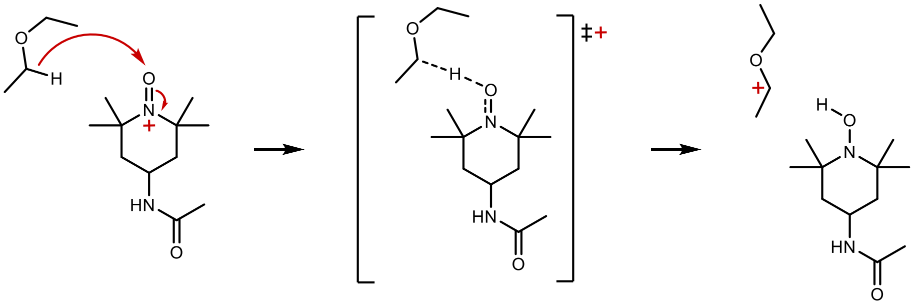
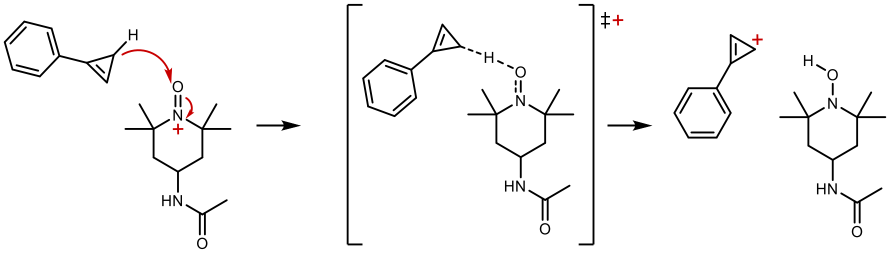
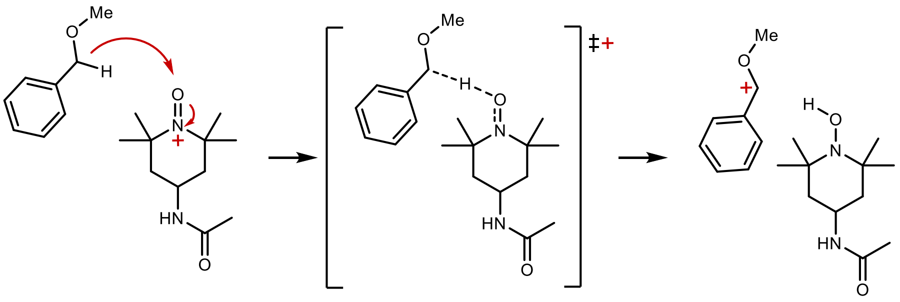
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
75		3.66	6.58	-2.91
76		13.85	18.38	-4.54
77		12.93	5.40	7.53
78		13.54	29.68	-16.14
79		18.60	11.78	6.82
80		1.14	15.53	-14.39
81		-0.11	14.46	-14.57
82		16.39	25.48	-9.09
83		15.83	25.03	-9.20

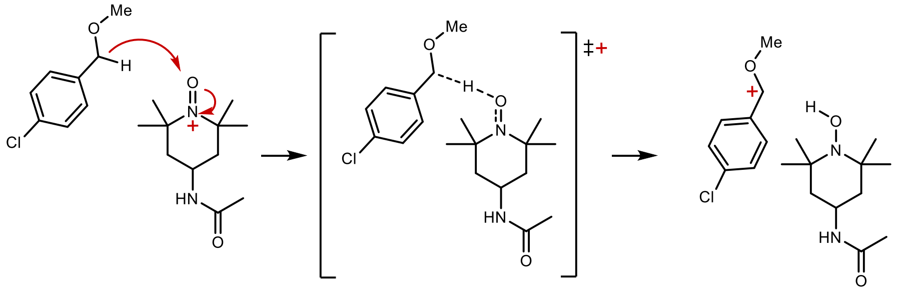
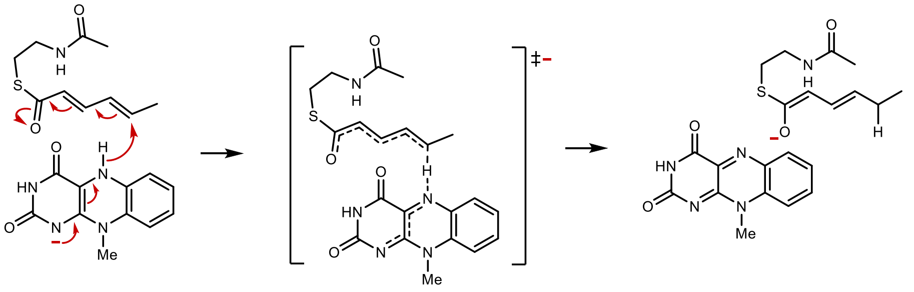
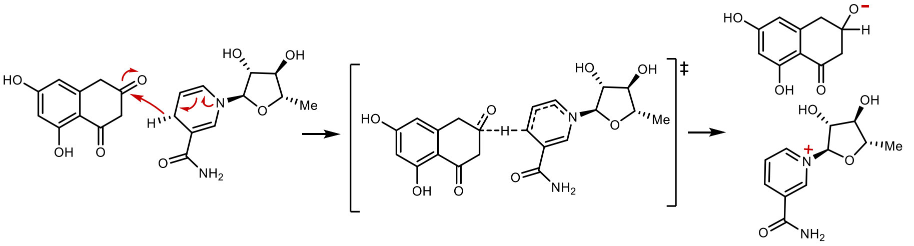
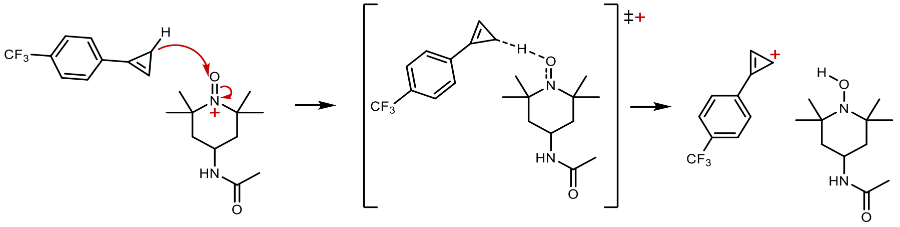
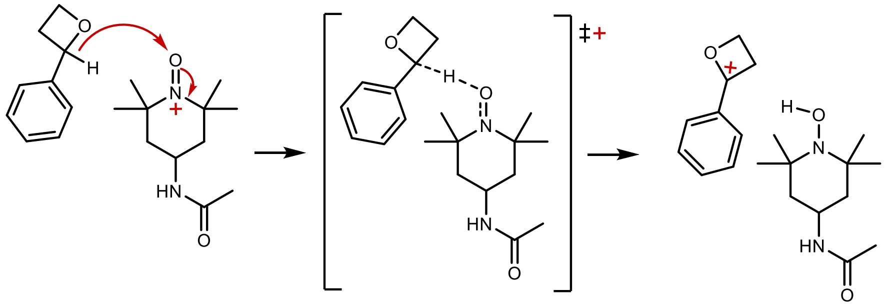
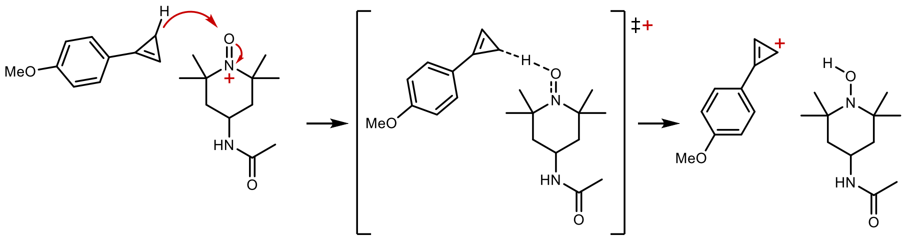
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
84		11.86	18.34	-6.48
85		15.58	27.22	-11.63
86		15.31	20.29	-4.98
87		3.84	18.14	-14.29
88		33.51	40.20	-6.69
89		13.47	22.19	-8.72
90		0.28	28.62	-28.35

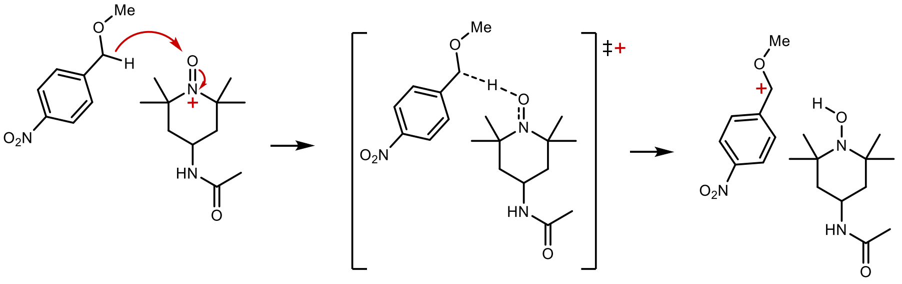
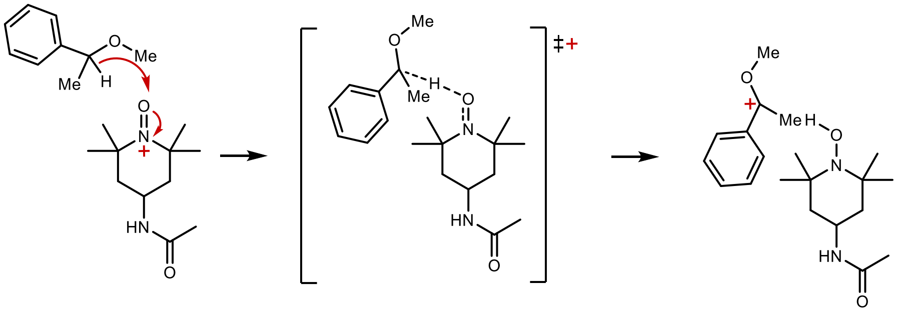
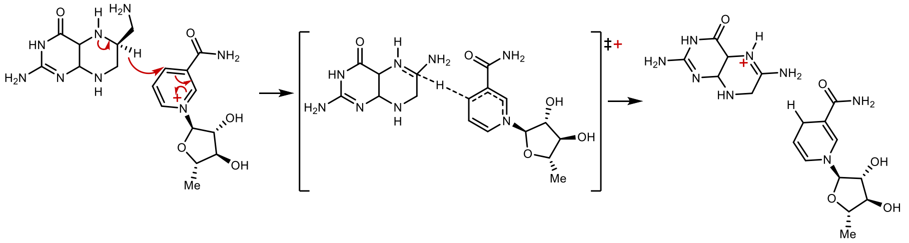
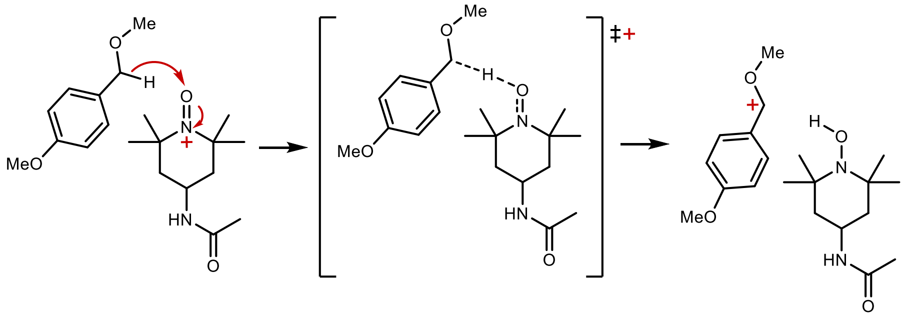
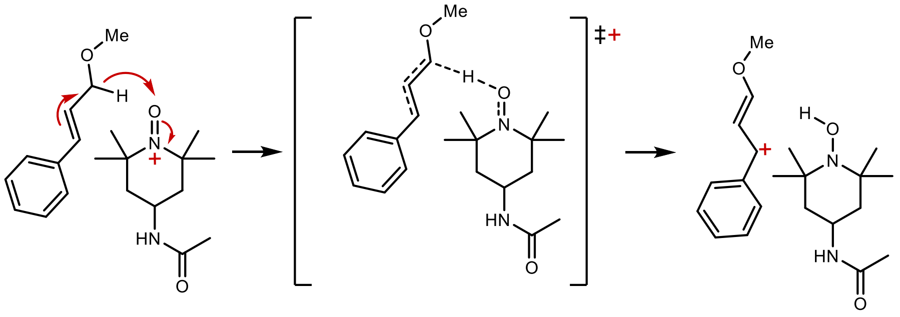
[a] units are in kcal/mol.

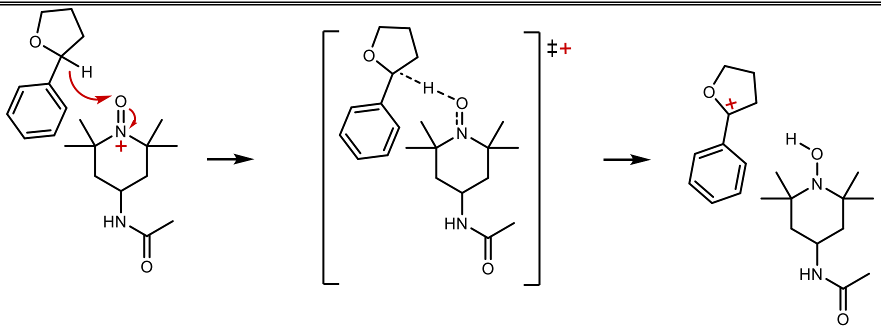
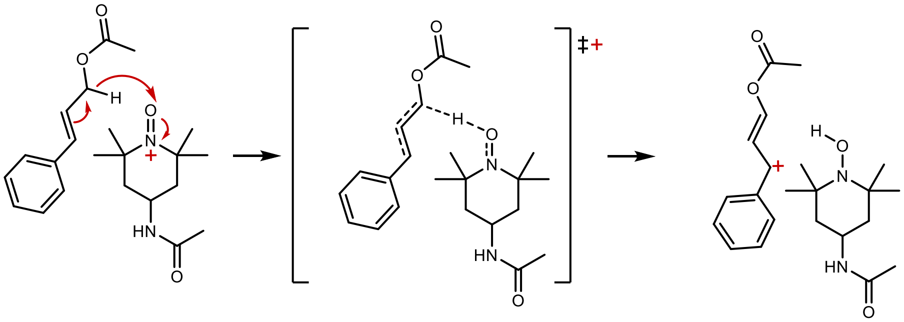
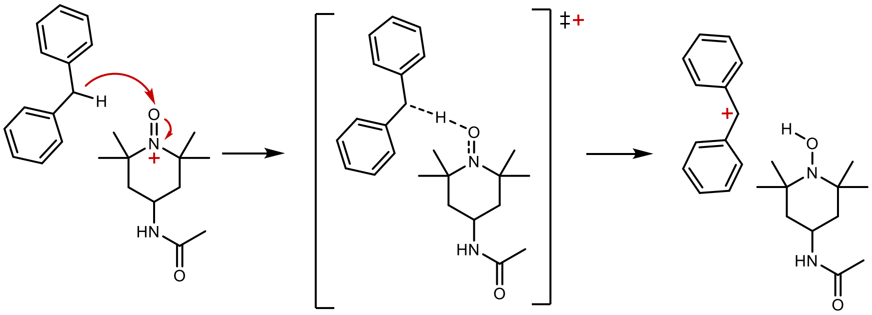
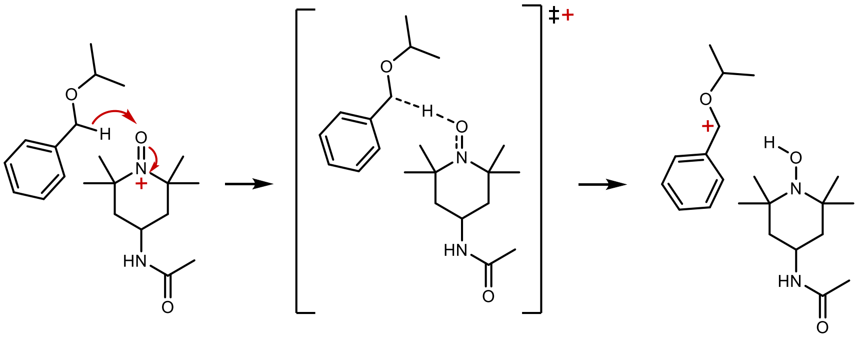
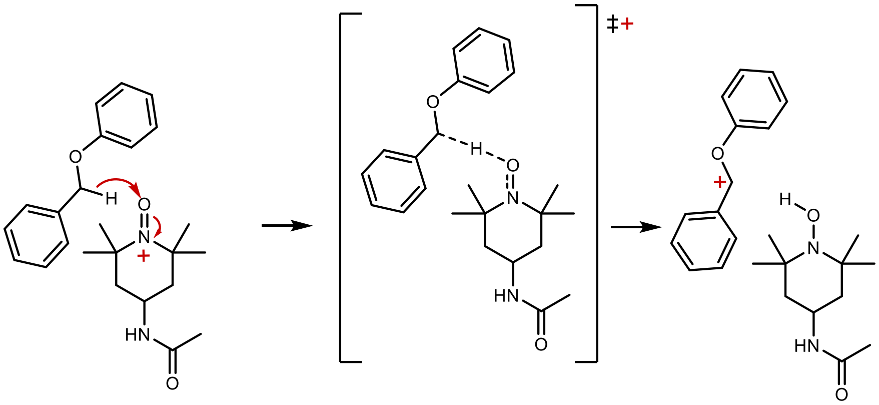
V. Hydride transfer

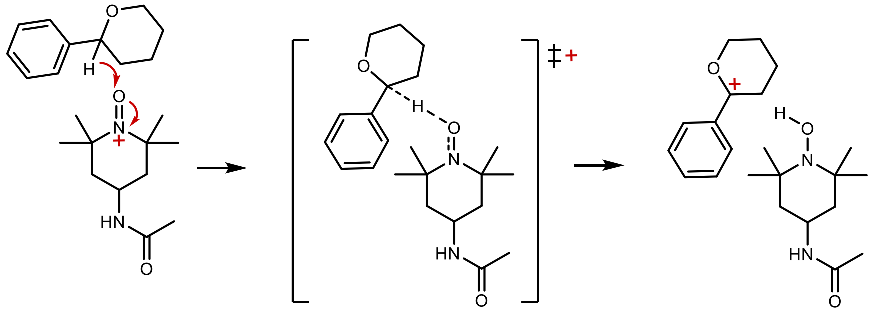
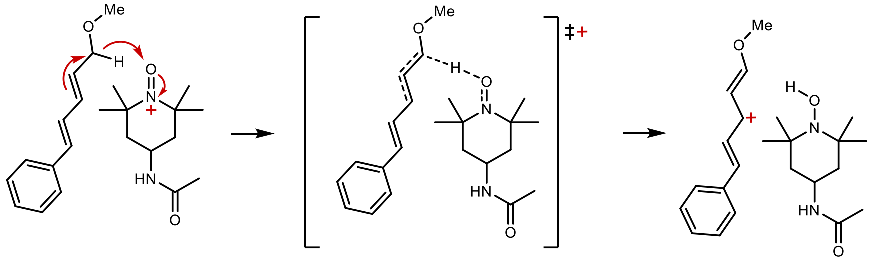
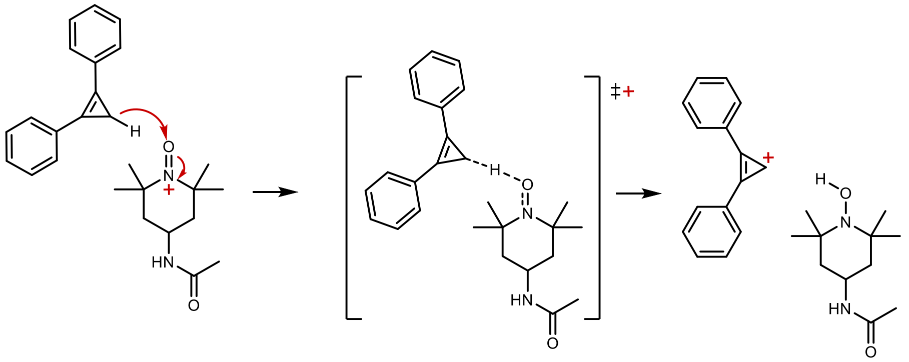
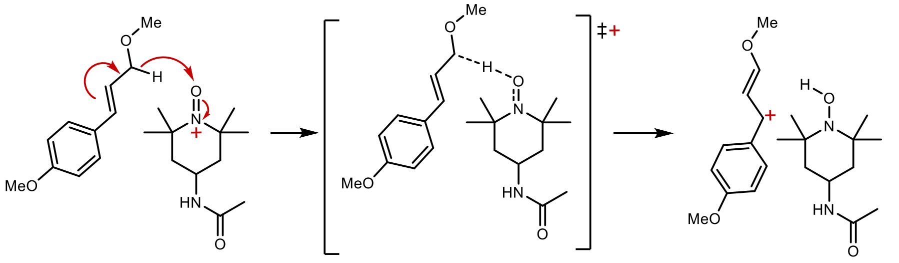
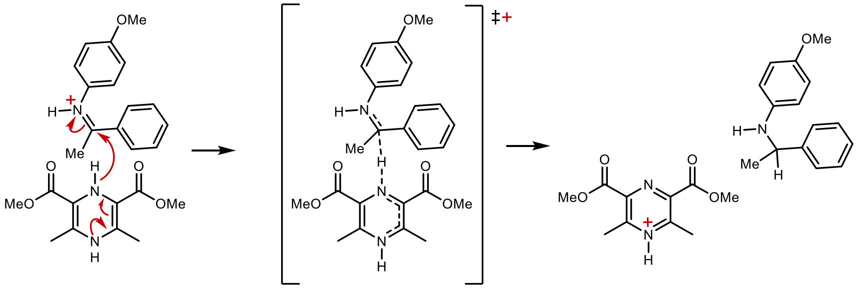
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
1		0.85	6.50	-5.65
2		22.01	23.81	-1.79
3		14.75	18.57	-3.82
4		5.49	25.25	-19.76
5		1.51	15.07	-13.56
6		10.00	11.94	-1.94
7		16.81	-0.50	17.31
8		1.04	8.07	-7.03
9		-2.46	8.58	-11.04

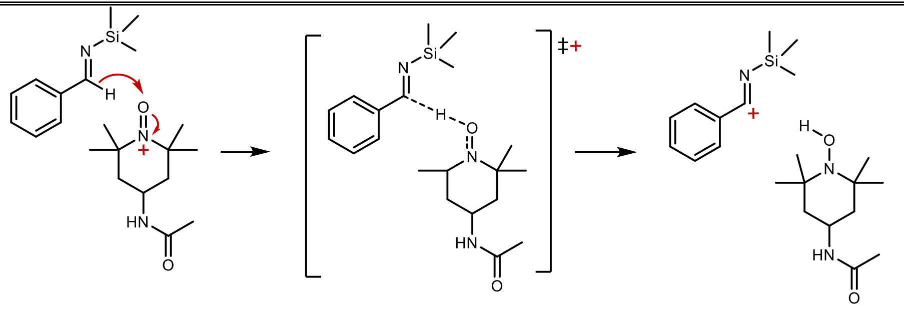
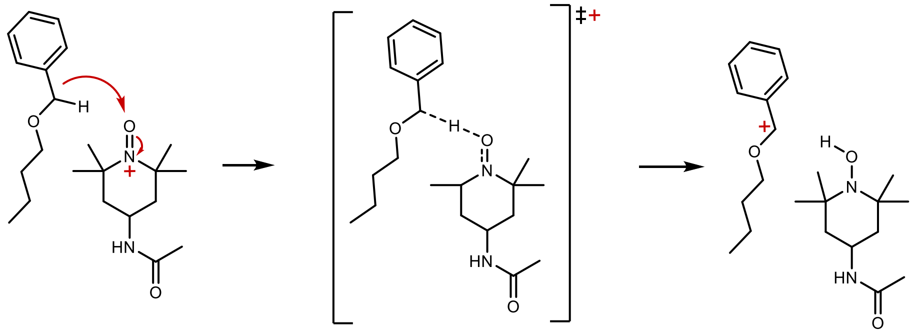
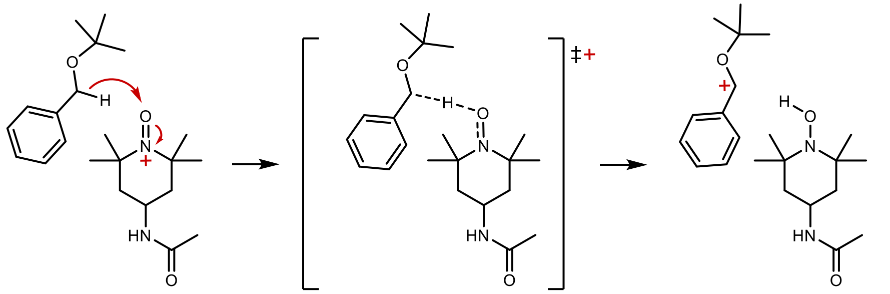
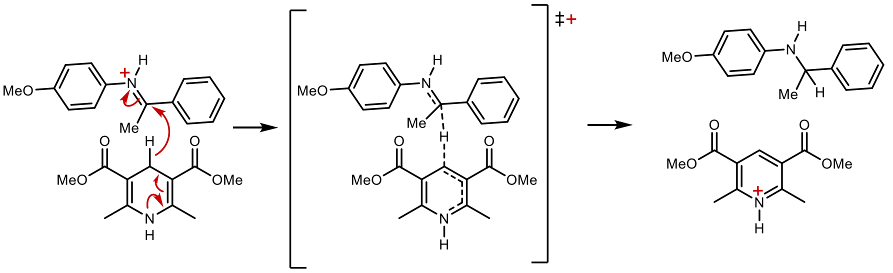
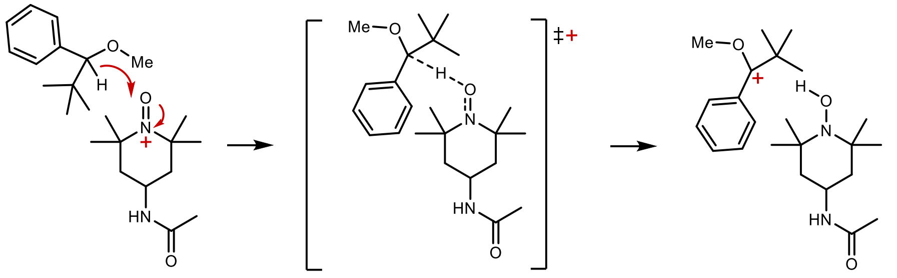
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
10		14.88	5.63	9.25
11		14.99	-4.21	19.20
12		5.79	20.17	-14.38
13		17.06	10.36	6.70
14		7.58	16.78	-9.21
15		13.40	18.28	-4.88

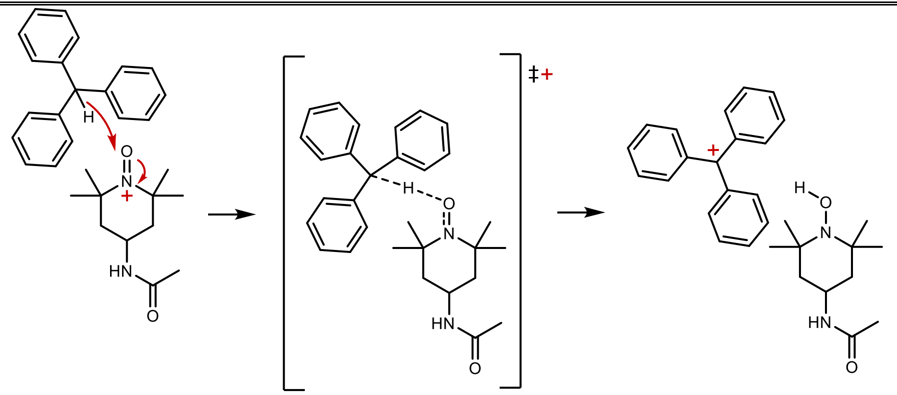
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
16		16.06	18.89	-2.83
17		14.02	3.43	10.58
18		20.63	-96.26	116.88
19		12.54	14.68	-2.14
20		5.38	16.10	-10.72
21		3.60	19.57	-15.96

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
22		21.77	13.81	7.95
23		12.53	21.75	-9.23
24		14.19	5.33	8.86
25		9.50	23.84	-14.33
26		4.49	17.53	-13.03

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
27		1.54	18.71	-17.17
28		9.84	12.24	-2.39
29		12.06	8.00	4.06
30		8.70	19.21	-10.52
31		15.27	18.26	-2.99

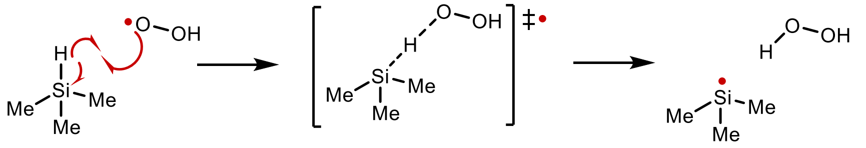
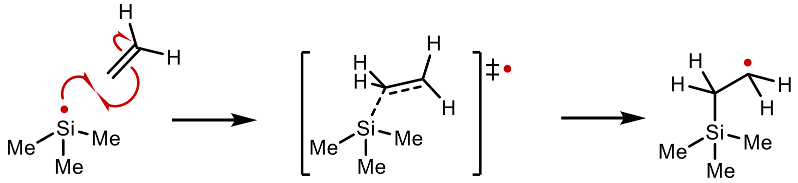
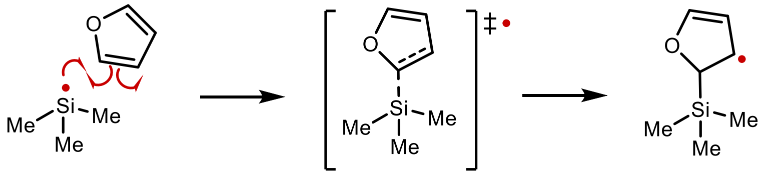
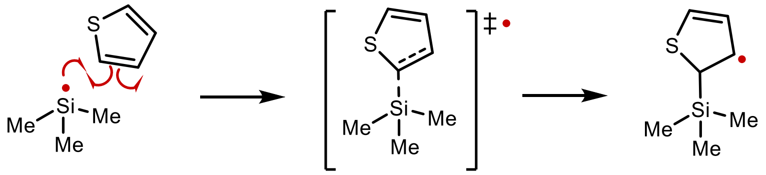
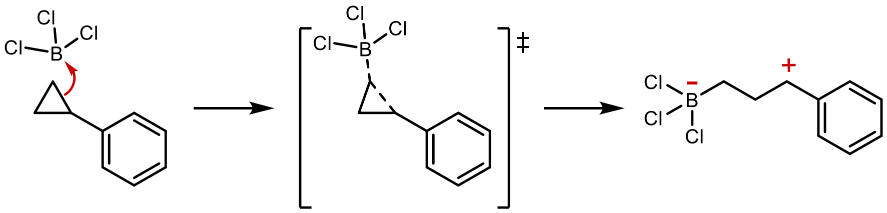
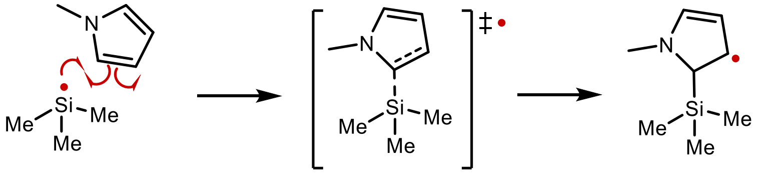
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
32		4.56	19.64	-15.08
33		1.97	19.13	-17.16
34		1.14	20.97	-19.83
35		0.74	20.87	-20.12
36		10.15	10.19	-0.04

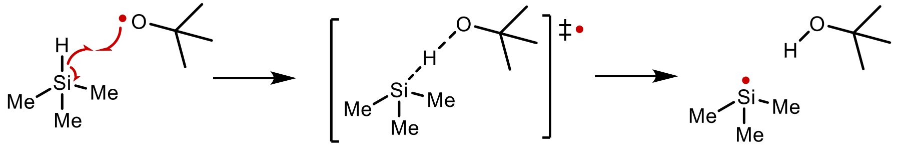
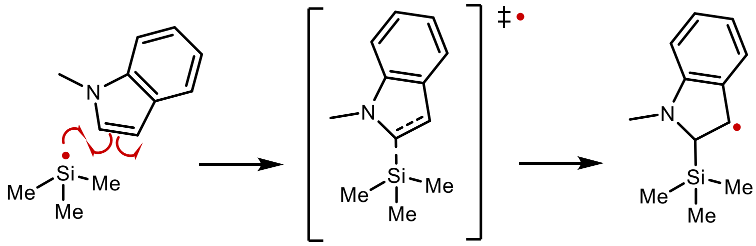
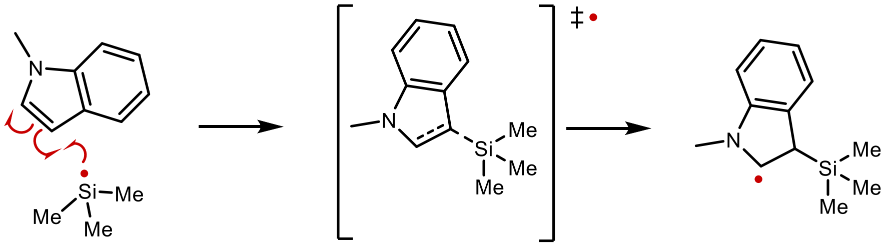
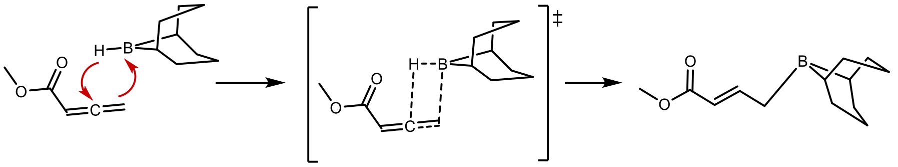
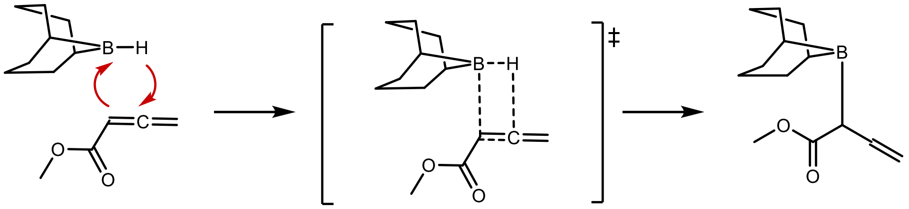
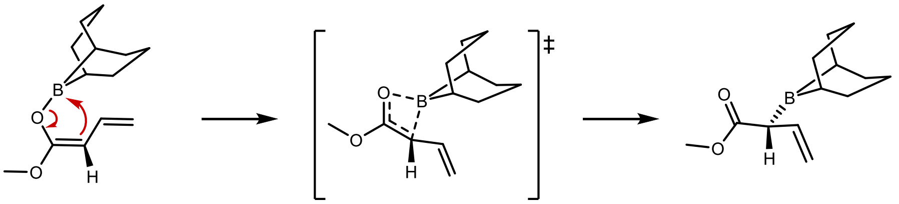
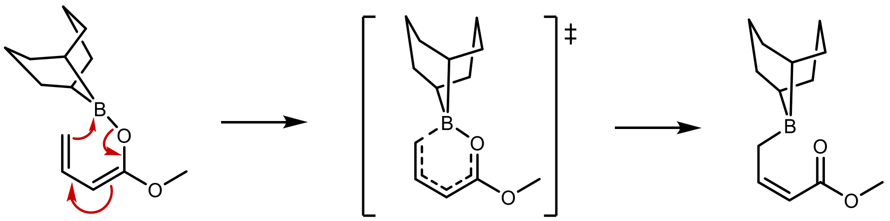
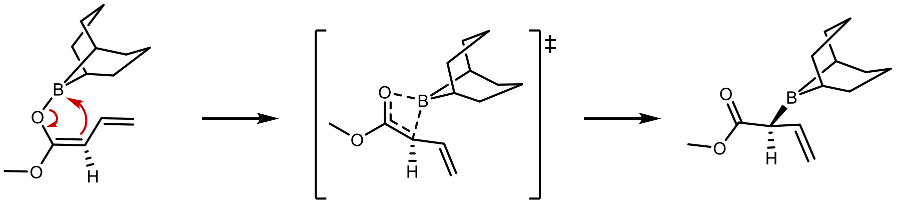
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
37		-0.57	30.56	-31.13
38		8.26	17.27	-9.02
39		5.34	18.84	-13.50
40		-1.16	7.89	-9.05
41		15.49	19.55	-4.06

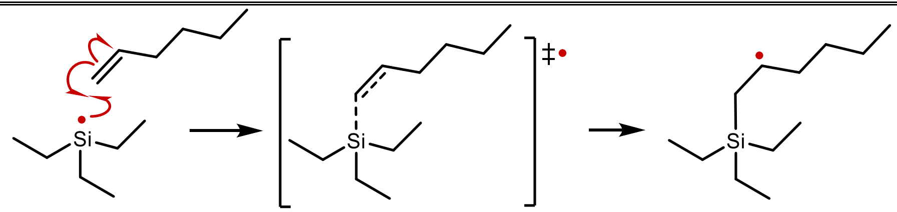
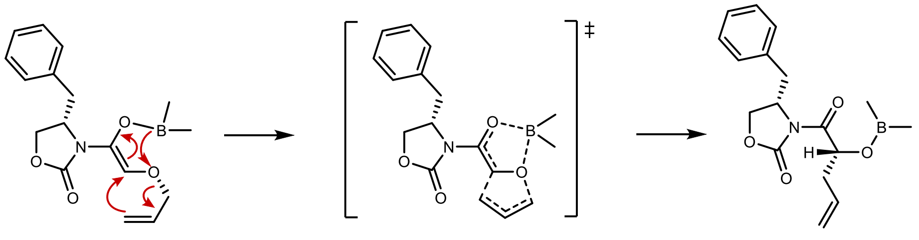
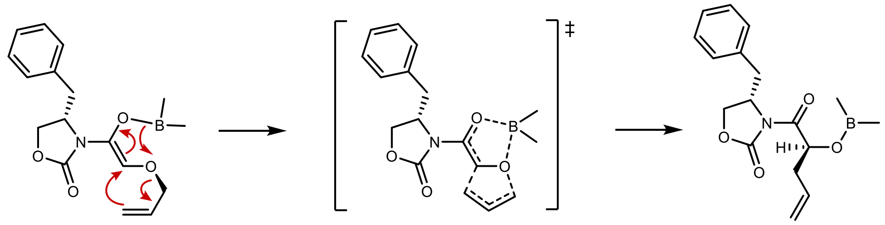
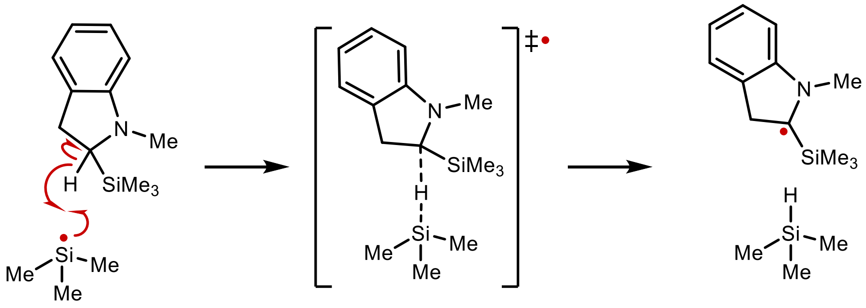
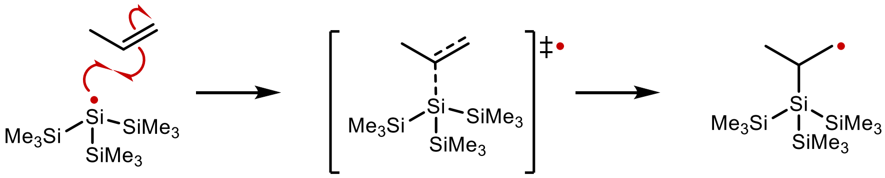
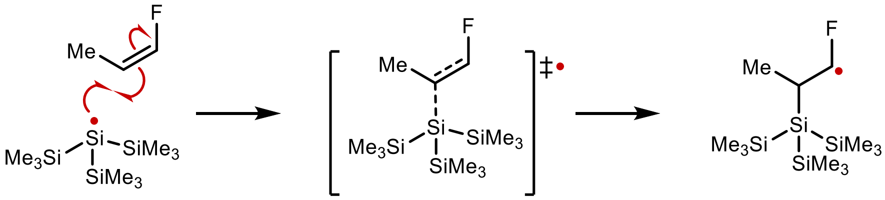
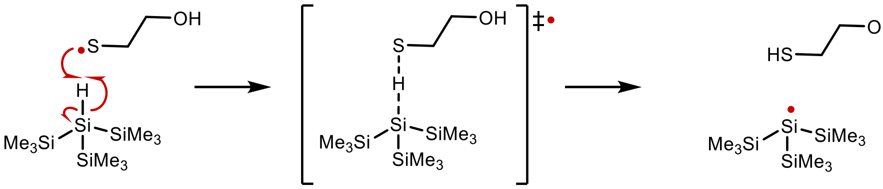
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
42		7.36	18.18	-10.82

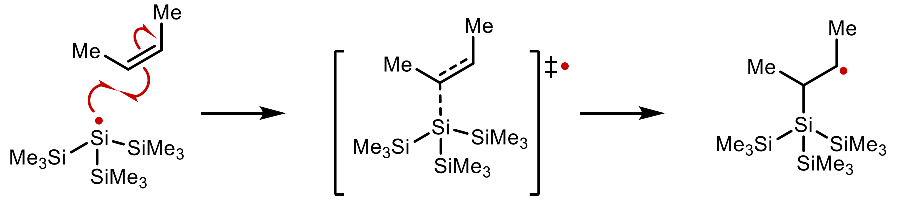
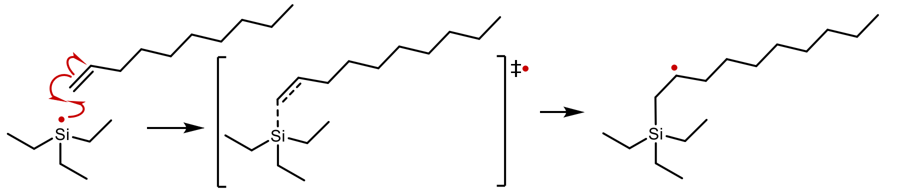
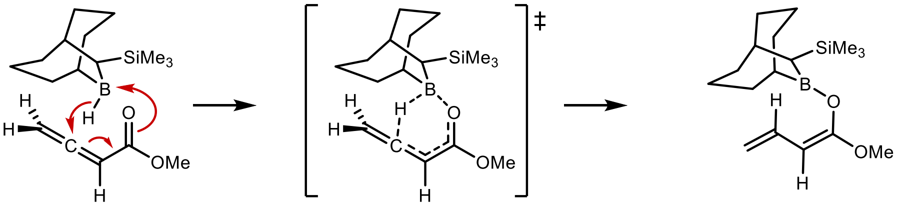
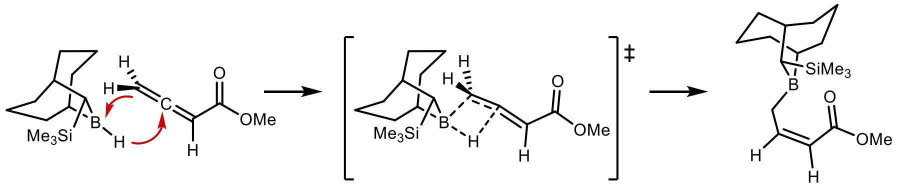
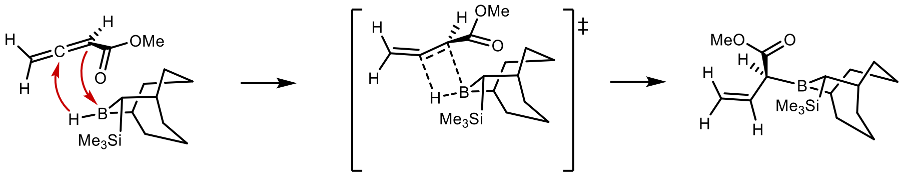
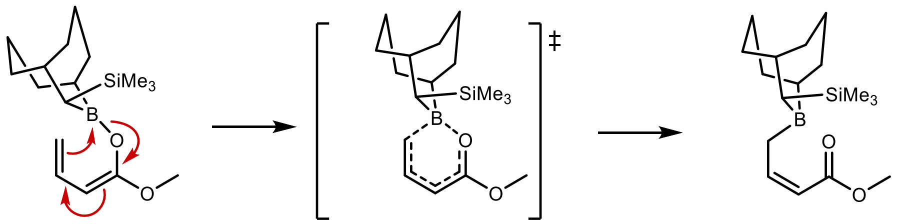
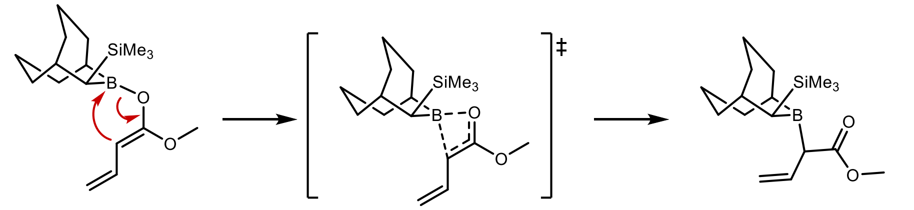
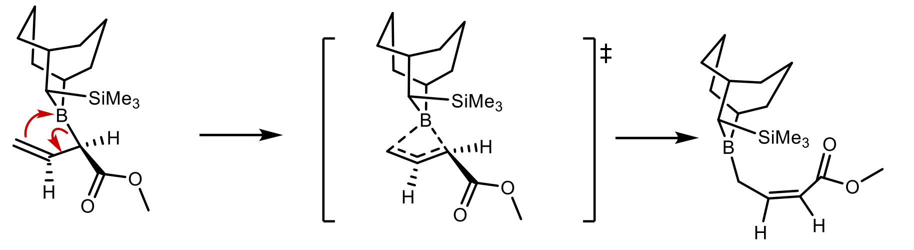
[a] units are in kcal/mol.

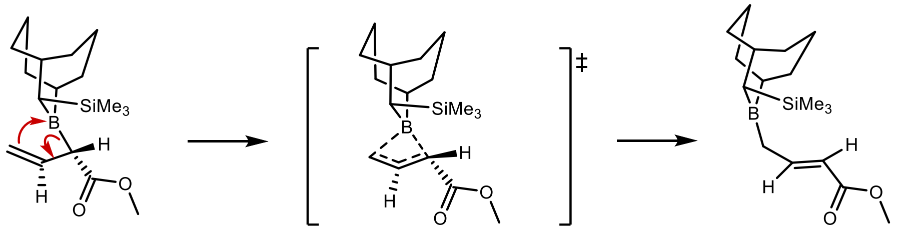
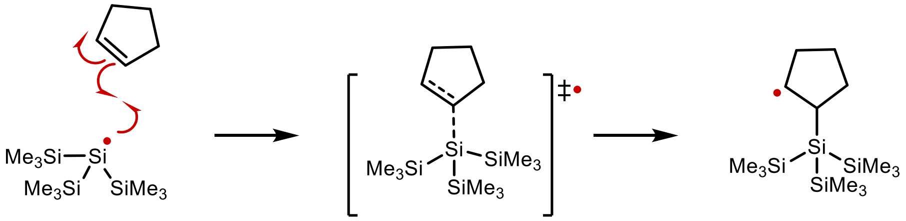
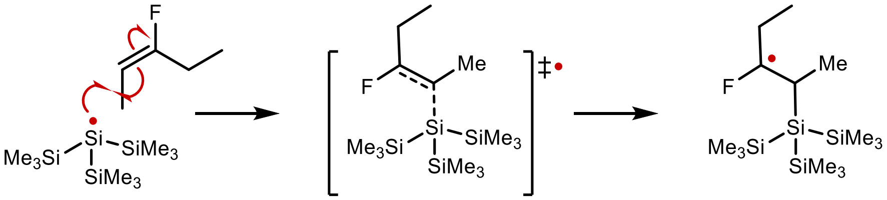
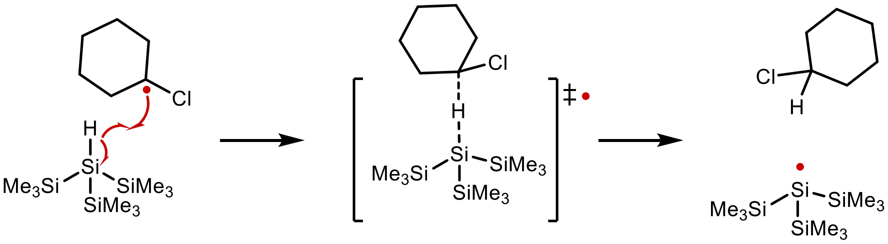
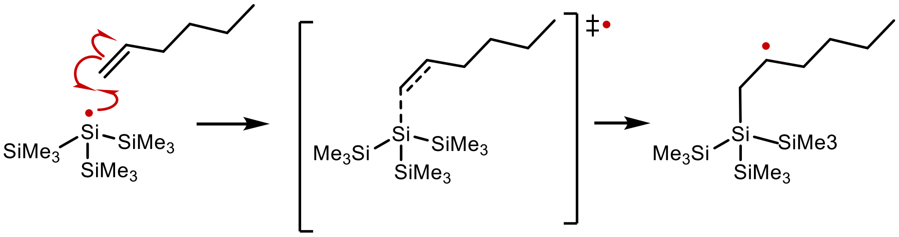
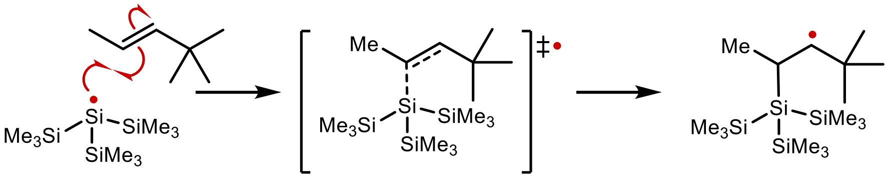
VI. B- and Si-containing

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
1		15.72	10.68	5.04
2		2.24	28.35	-26.12
3		4.81	25.87	-21.06
4		4.62	29.54	-24.91
5		14.75	-1.06	15.80
6		4.04	19.53	-15.48

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
7		3.89	19.12	-15.24
8		3.36	21.63	-18.27
9		5.78	21.11	-15.32
10		3.49	52.87	-49.38
11		1.31	50.23	-48.92
12		20.32	14.77	5.54
13		11.81	16.75	-4.95
14		13.62	10.81	2.80

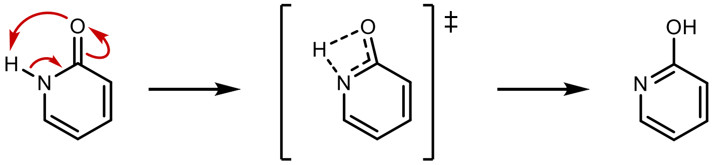
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
15		1.33	27.06	-25.73
16		18.86	53.61	-34.76
17		20.99	54.73	-33.74
18		7.60	15.41	-7.82
19		4.39	13.75	-9.36
20		4.00	14.18	-10.18
21		-1.69	2.54	-4.23

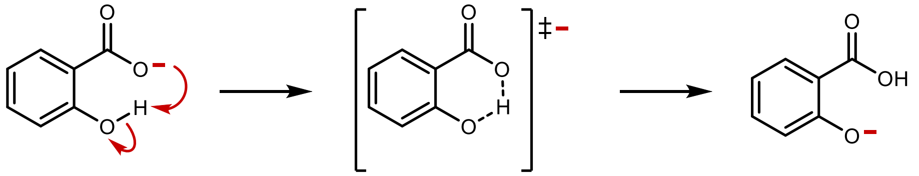
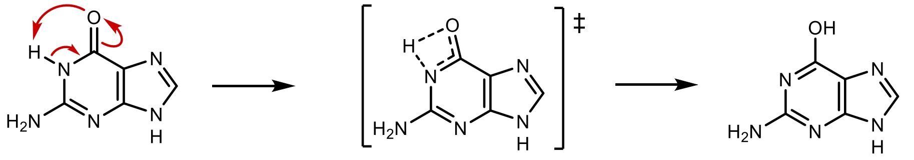
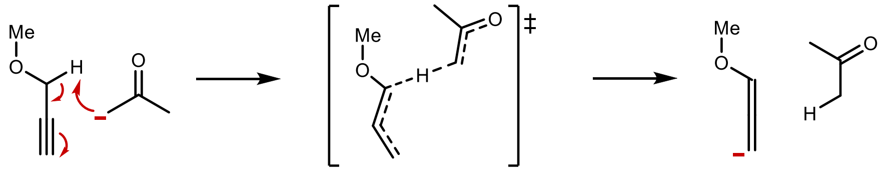
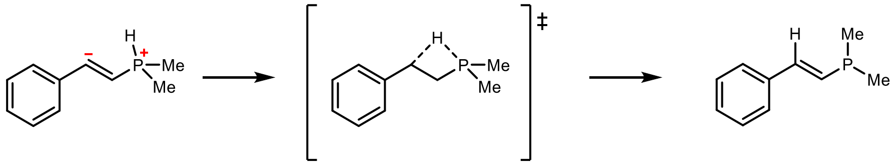
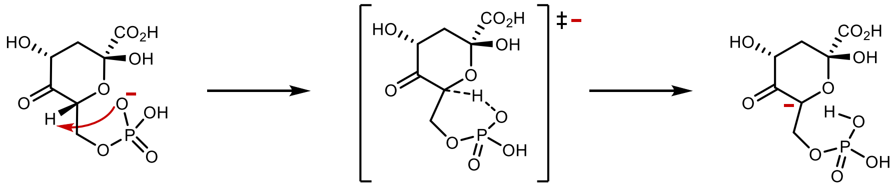
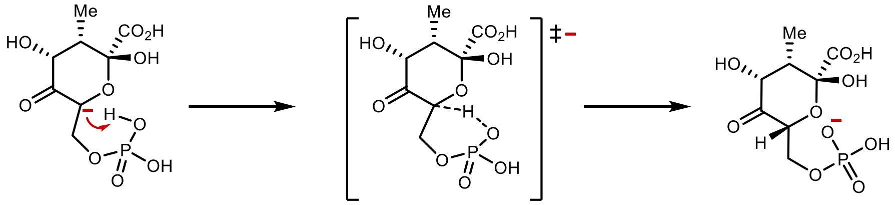
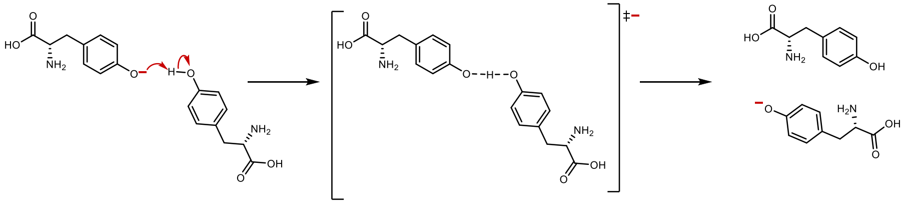
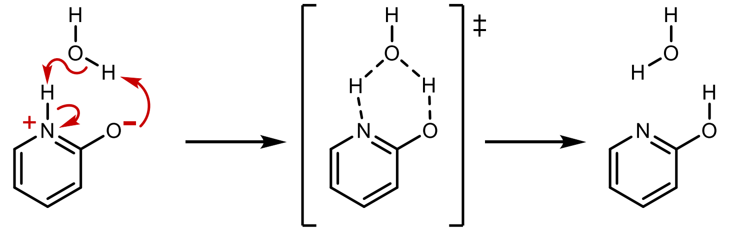
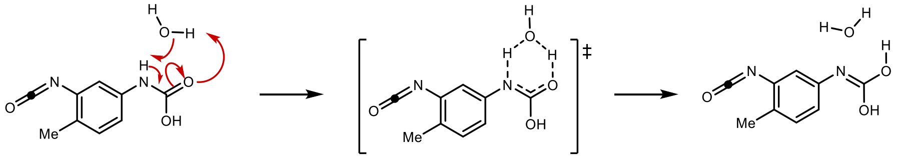
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
22		2.98	12.74	-9.76
23		1.07	26.52	-25.45
24		7.84	61.88	-54.04
25		8.11	56.40	-48.28
26		8.68	55.38	-46.70
27		21.41	15.51	5.90
28		20.08	15.35	4.73
29		18.58	20.49	-1.91

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
30		13.88	15.61	-1.73
31		2.21	13.33	-11.12
32		2.61	13.62	-11.00
33		1.43	17.74	-16.31
34		1.41	13.67	-12.26
35		5.79	13.34	-7.55

[a] units are in kcal/mol.

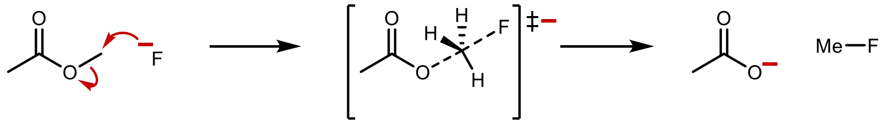
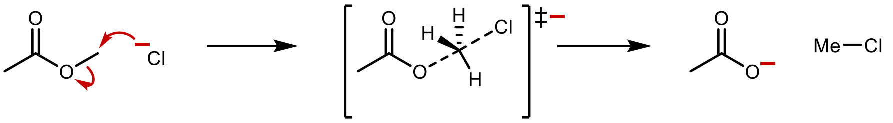
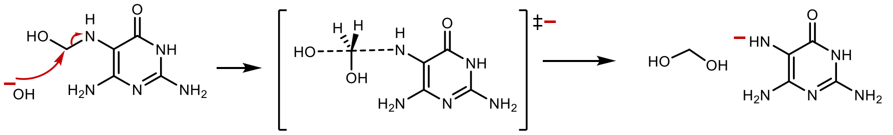
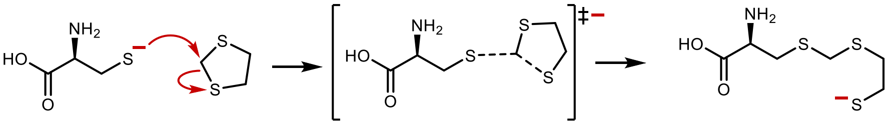
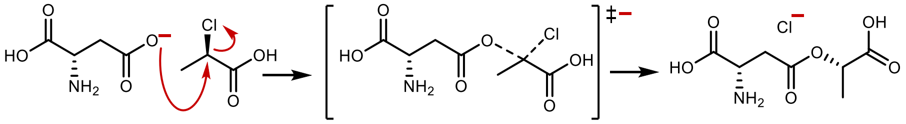
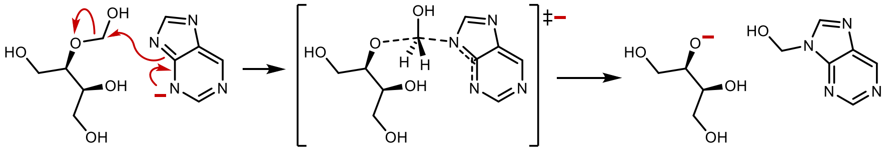
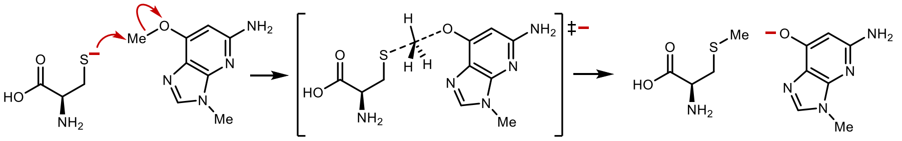
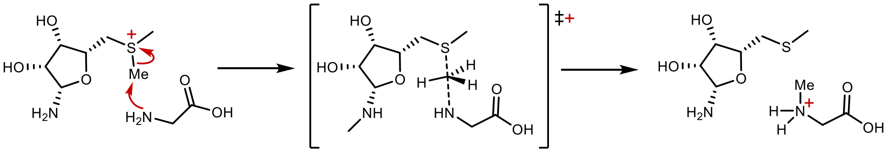
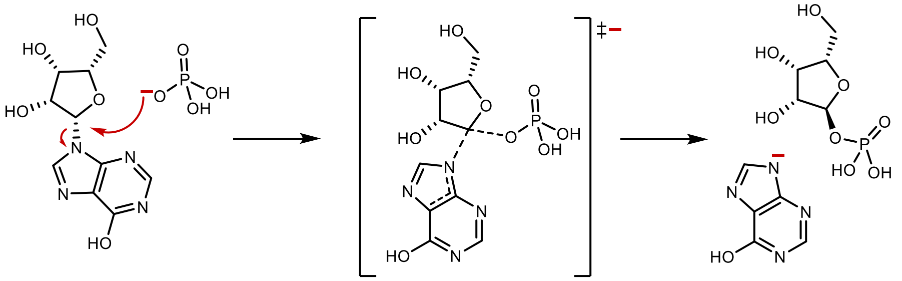
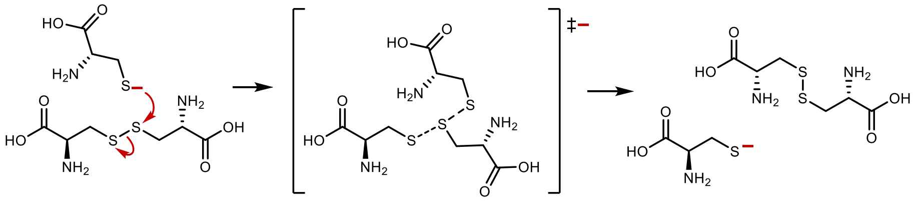
VII. Proton transfer

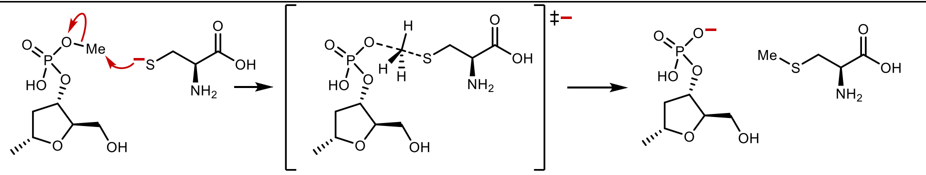
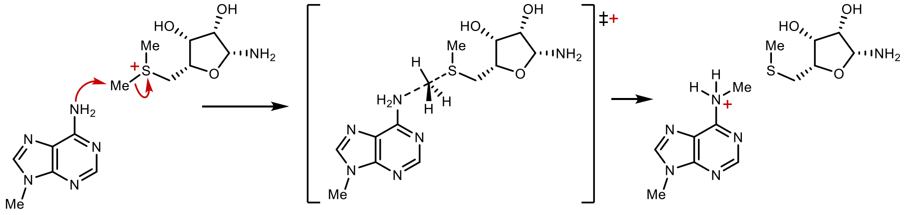
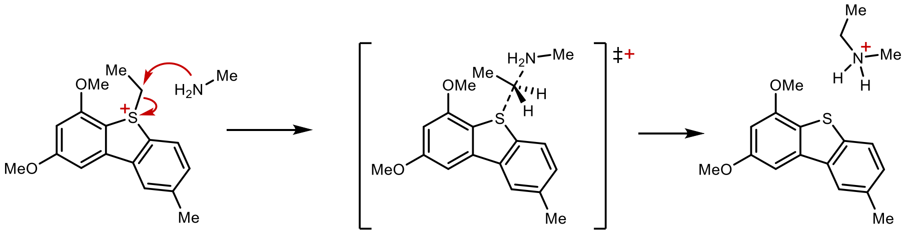
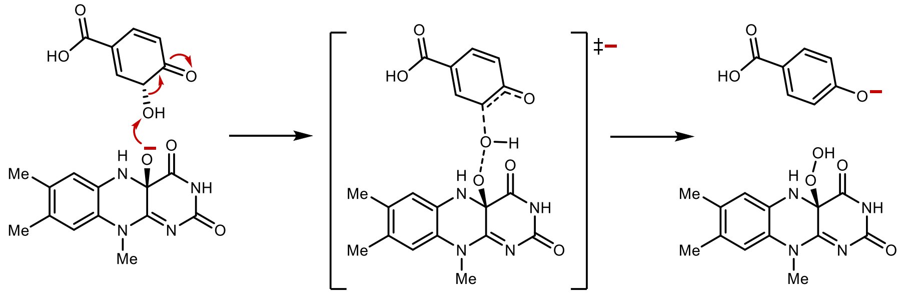
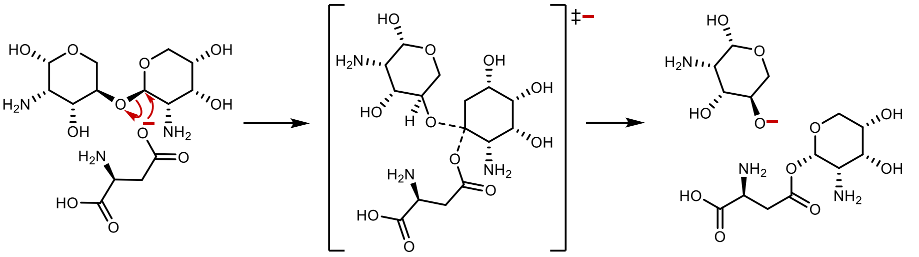
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
1		37.29	38.20	-0.91

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
2		0.84	0.75	0.09
3		37.38	37.14	0.24
4		12.73	0.08	12.64
5		11.94	71.12	-59.18
6		30.60	3.14	27.46
7		3.08	32.75	-29.68
8		-31.63	-30.51	-1.12
9		4.73	5.64	-0.91
10		17.32	-0.47	17.79

[a)] units are in kcal/mol.

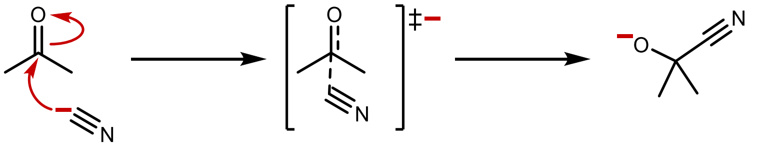
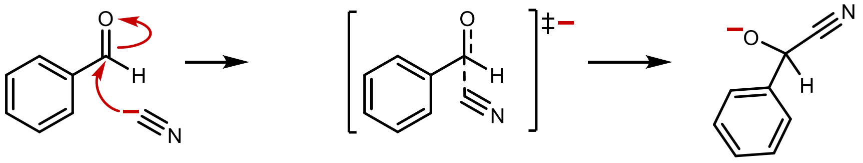
VIII. Nucleophilic substitution

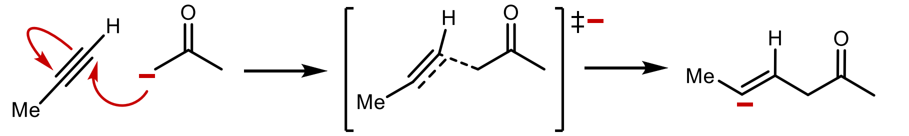
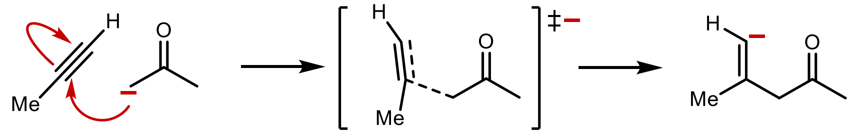
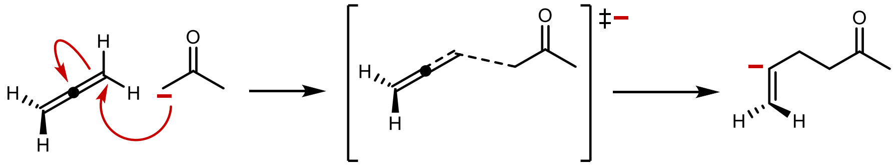
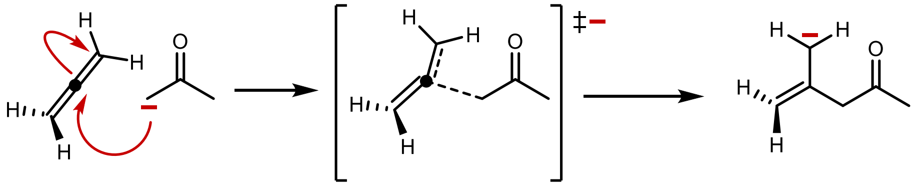
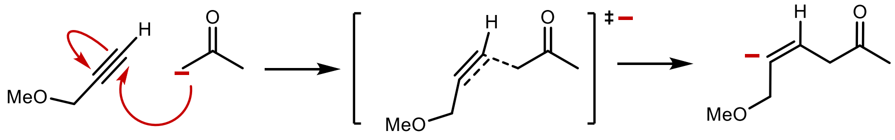
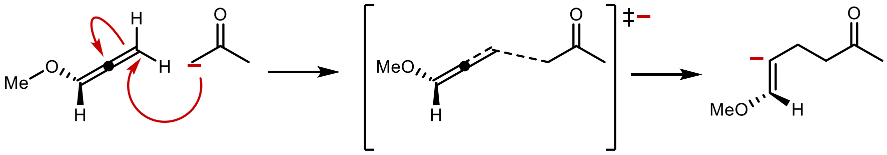
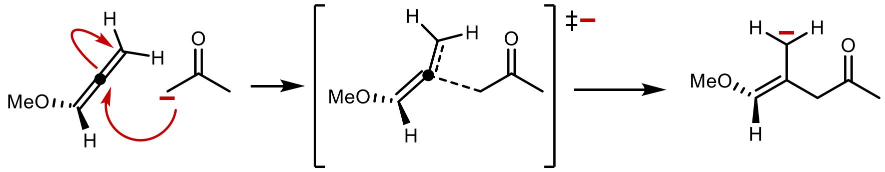
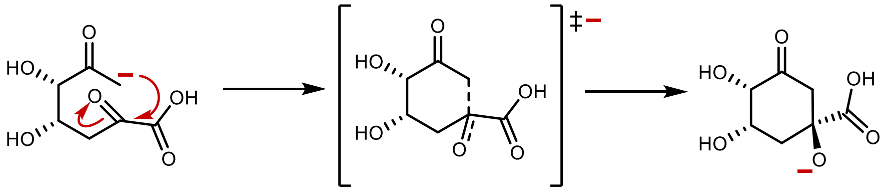
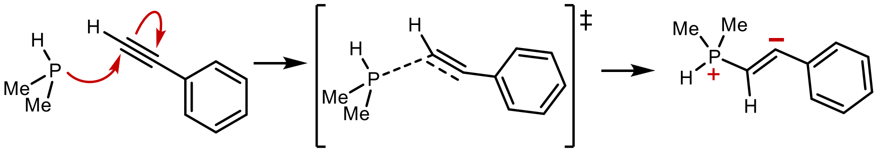
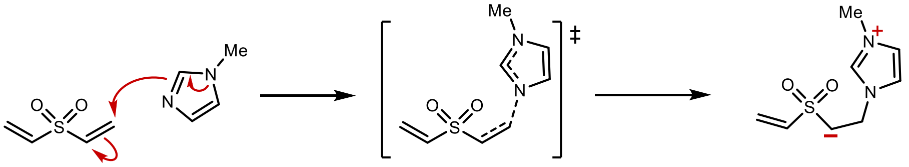
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
1		-3.21	15.07	-18.28
2		14.28	-0.05	14.33
3		13.63	28.04	-14.41
4		23.74	29.96	-6.22
5		-4.00	4.56	-8.56
6		20.98	10.61	10.37
7		9.10	11.86	-2.76
8		17.89	15.80	2.09
9		3.08	-5.70	8.78
10		3.14	6.09	-2.95

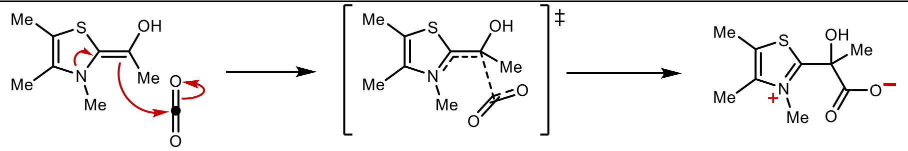
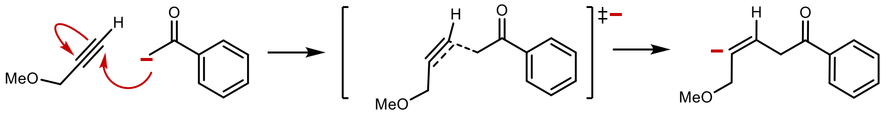
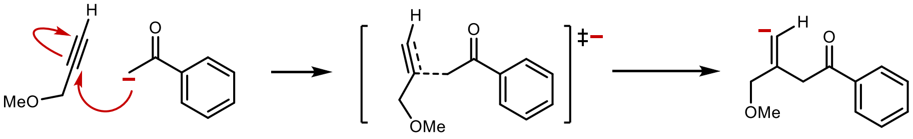
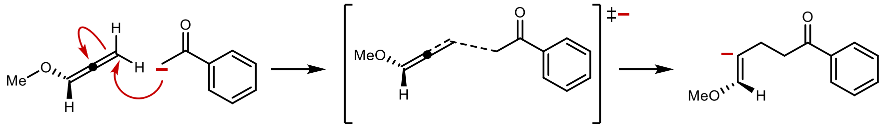
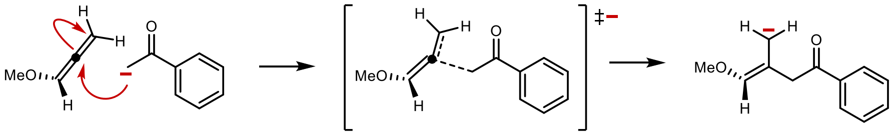
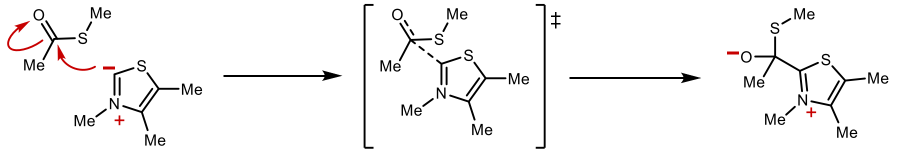
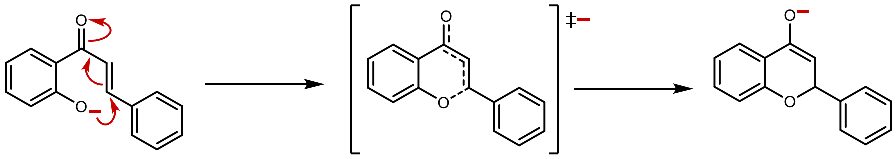
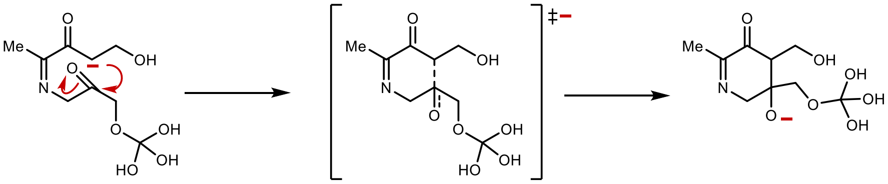
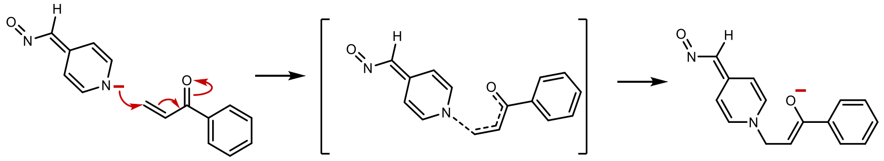
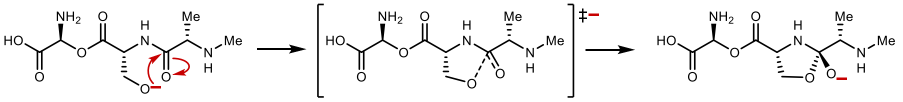
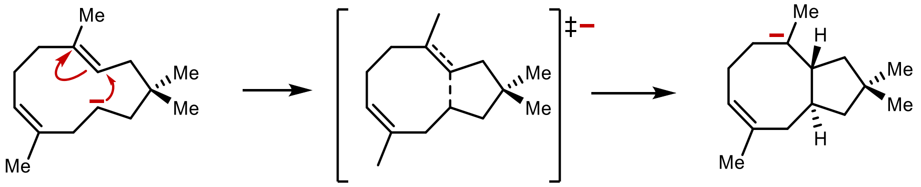
#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
11		-2.99	19.74	-22.72
12		14.76	6.60	8.16
13		12.46	18.94	-6.48
14		32.38	-2.33	34.71
15		18.15	2.87	15.27

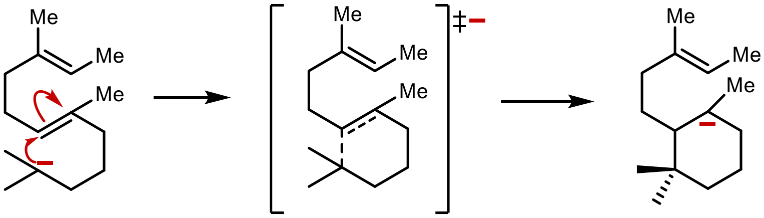
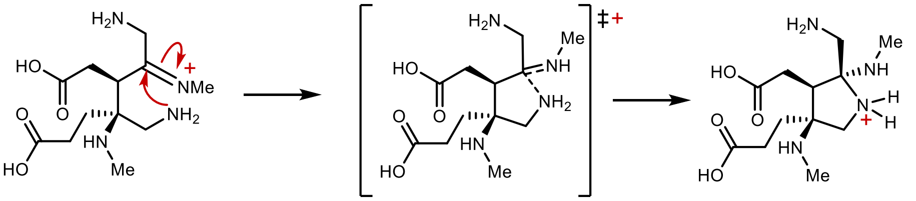
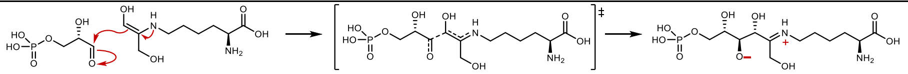
[a] units are in kcal/mol.

IX. Nucleophilic addition

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
1		-0.94	4.81	-5.76
2		-4.35	6.35	-10.71

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
3		20.11	29.26	-9.15
4		17.01	24.61	-7.59
5		18.75	26.00	-7.24
6		7.22	32.29	-25.06
7		2.72	17.07	-14.35
8		1.90	15.44	-13.54
9		7.30	33.00	-25.70
10		0.22	7.33	-7.11
11		31.22	14.12	17.10
12		13.77	0.34	13.43

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
13		7.18	9.50	-2.31
14		4.89	13.69	-8.80
15		5.42	8.25	-2.83
16		3.82	12.73	-8.92
17		9.95	30.33	-20.38
18		6.92	24.11	-17.19
19		20.06	9.65	10.41
20		1.46	11.03	-9.57
21		1.23	11.93	-10.69
22		12.21	3.51	8.69
23		6.35	13.38	-7.03

#	Mechanism	Forward BH[a]	Reverse BH[a]	RE[a]
24		7.58	22.45	-14.87
25	 H2N HO O NMe NH2 HN Me ‡ NH N H	3.76	8.69	-4.93
26		6.41	5.53	0.88

[a]) units are in kcal/mol.

Appendix 4

Supporting Information for Chapter 7

Section S1. Formulas of statistical error measures

j) Mean absolute error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where, $x_i = |x_{calc,i} - x_{ref,i}|$

k) Mean signed error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where, $x_i = x_{calc,i} - x_{ref,i}$

l) Maximum absolute error (MAXE)

$$MAXE = \max_i |x_{calc,i} - x_{ref,i}|$$

m) Root-mean-square error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

where, $x_i = x_{calc,i} - x_{ref,i}$

n) Standard deviation (SD)

$$SD = \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

where, $x_i = x_{calc,i} - x_{ref,i}$

$$\overline{x} = \frac{1}{n}(x_{calc,i} - x_{ref,i})$$

Section S2. Sample input file demonstrating the use of atom-centered potentials in Gaussian software

The MINIX and 6-31G* basis set files (in .gbs extension) and the corresponding ACP files (in .acp extension) are provided separately in the supporting information ZIP file accompanying this document.

An externally specified basis set file named “*minix.gbs*” and the additional ACP file “*minix.acp*” is defined and invoked by adding the keyword “*genECP*” to the route section of the Gaussian input file. Note that the ACPs are not transferable and are proposed to be used with their underlying methods only.

```
%mem=4GB
%nprocs=8
# HF genECP

Title: Sample water dimer input using HF/MINIX-ACP method

0 1
O   -0.702196054   -0.056060256    0.009942262
H   -1.022193224    0.846775782   -0.011488714
H    0.257521062    0.042121496    0.005218999
O    2.220871067    0.026716792    0.000620476
H    2.597492682   -0.411663274    0.766744858
H    2.593135384   -0.449496183   -0.744782026

@minix.gbs/N

@minix.acp/N
```

Section S3. Basis set file for MINIX

```
-H     0
S   3   1.00
      7.034063              0.070452
      1.064756              0.407826
      0.236559              0.647752
****
-B     0
S   3   1.00
      4.457854             -0.082419
      0.369315              0.559064
      0.122555              0.516795
S   3   1.00
    108.43704               0.068651
     16.120560              0.389933
      3.3734300             0.671395
P   3   1.00
      3.214892              0.105900
      0.646136              0.457180
      0.153916              0.631861
****
-C     0
S   3   1.00
      6.616612             -0.081380
      0.525856              0.574853
      0.169958              0.502413
S   3   1.00
    153.17226               0.070740
     23.073030              0.395380
      4.9232900             0.663311
P   3   1.00
      4.912920              0.109931
      0.997616              0.462713
      0.232685              0.627514
****
-N     0
S   3   1.00
```

```
      8.919426            -0.080890
      0.706141             0.567202
      0.225054             0.511092
S   3   1.00
    218.36449              0.067870
     32.598890             0.390202
      6.9173900            0.670083
P   3   1.00
      6.556272             0.115919
      1.349079             0.469958
      0.312209             0.618448
****
-O     0
S   3   1.00
     11.789326            -0.080820
      0.9128940            0.582090
      0.2866610            0.497160
S   3   1.00
    281.86658              0.069060
     42.416000             0.393159
      9.0956200            0.665669
P   3   1.00
      8.274140             0.124271
      1.715463             0.476594
      0.383013             0.613044
****
-F     0
S   3   1.00
    368.37112              0.067040
     55.061060             0.389249
     11.747670             0.670788
S   3   1.00
     15.364708            -0.080550
      1.1675460            0.587729
      0.3631410            0.491979
P   3   1.00
     10.725667             0.126270
      2.2258170            0.477948
      0.4861050            0.614008
****
-Si     0
S   3   1.00
    909.2348700            0.0664050
    137.1245600            0.3862220
     29.7148100            0.6722400
S   3   1.00
     39.1294230           -0.0909990
      3.3359810            0.6116150
      1.2512590            0.4568600
S   3   1.00
      2.1976490           -0.1687330
      0.2759270            0.6754530
      0.1004250            0.4294190
P   3   1.00
     37.8817610            0.1087530
      8.3045980            0.4635150
      2.1207920            0.6113340
P   3   1.00
      0.5457890            0.2389130
      0.2082200            0.5422950
      0.0760070            0.3454530
D   1   1.00
      0.3500000            1.0000000
****
```

```
-P     0
S   3   1.00
   1053.2658000             0.0658650
    158.7904400             0.3845780
     34.4244070             0.6739630
S   3   1.00
     45.4503770            -0.0926550
      3.8999260             0.6265130
      1.4885070             0.4410390
S   3   1.00
      2.4694830            -0.1805490
      0.3208720             0.6809520
      0.1168320             0.4291420
P   3   1.00
     46.1000190             0.1053880
     10.1650570             0.4597120
      2.6447940             0.6137140
P   3   1.00
      0.6790590             0.2358850
      0.2578260             0.5541600
      0.0927830             0.3365300
D   1   1.00
      0.4500000             1.0000000
****
-S     0
S   3   1.00
   1201.4584000             0.0657650
    181.3921200             0.3839480
     39.4047950             0.6743720
S   3   1.00
     52.1390300            -0.0942320
      4.5287990             0.6354680
      1.7549380             0.4315060
S   3   1.00
      2.9205260             0.1900420
      0.3921870            -0.6855270
      0.1426990            -0.4292720
P   3   1.00
     54.6440710             0.1036730
     12.1229020             0.4581900
      3.2065040             0.6134000
P   3   1.00
      0.8876150             0.2294360
      0.1117430             0.3537000
      0.3271000             0.5529600
D   1   1.00
      0.5500000             1.0000000
****
-Cl     0
S   3   1.00
   1362.0220000             0.0655440
    205.8111000             0.3829870
     44.7721670             0.6752100
S   3   1.00
     59.2257320            -0.0956200
      5.2139020             0.6414260
      2.0473460             0.4251530
S   3   1.00
      3.4471240             0.1964010
      0.4737850            -0.6923600
      0.1713210            -0.4261930
P   3   1.00
     64.0999580             0.1017890
     14.2871390             0.4561070
     3.8281350            0.6142820
P   3   1.00
     1.1039040            0.2359030
     0.1332360            0.3466000
     0.3991780            0.5580660
D   1   1.00
     0.6500000            1.0000000
****
```

Section S4. Basis set file for 6-31G*

```
-H     0
S   3   1.00
    18.7311370            0.03349460
     2.8253937            0.23472695
     0.6401217            0.81375733
S   1   1.00
     0.1612778            1.0000000
****
-B     0
S   6   1.00
  2068.8823000            0.0018663
   310.6495700            0.0142515
    70.6830330            0.0695516
    19.8610800            0.2325729
     6.2993048            0.4670787
     2.1270270            0.3634314
SP   3   1.00
     4.7279710           -0.1303938            0.0745976
     1.1903377           -0.1307889            0.3078467
     0.3594117            1.1309444            0.7434568
SP   1   1.00
     0.1267512            1.0000000            1.0000000
D   1   1.00
     0.6000000            1.0000000
****
-C     0
S   6   1.00
  3047.5249000            0.0018347
   457.3695100            0.0140373
   103.9486900            0.0688426
    29.2101550            0.2321844
     9.2866630            0.4679413
     3.1639270            0.3623120
SP   3   1.00
     7.8682724           -0.1193324            0.0689991
     1.8812885           -0.1608542            0.3164240
     0.5442493            1.1434564            0.7443083
SP   1   1.00
     0.1687144            1.0000000            1.0000000
D   1   1.00
     0.8000000            1.0000000
****
-N     0
S   6   1.00
  4173.5110000            0.0018348
   627.4579000            0.0139950
   142.9021000            0.0685870
    40.2343300            0.2322410
    12.8202100            0.4690700
     4.3904370            0.3604550
SP   3   1.00
    11.6263580           -0.1149610            0.0675800
```

```
     2.7162800           -0.1691180            0.3239070
     0.7722180            1.1458520            0.7408950
SP   1   1.00
     0.2120313            1.0000000            1.0000000
D   1   1.00
     0.8000000            1.0000000
****
-O     0
S   6   1.00
  5484.6717000            0.0018311
   825.2349500            0.0139501
   188.0469600            0.0684451
    52.9645000            0.2327143
    16.8975700            0.4701930
     5.7996353            0.3585209
SP   3   1.00
    15.5396160           -0.1107775            0.0708743
     3.5999336           -0.1480263            0.3397528
     1.0137618            1.1307670            0.7271586
SP   1   1.00
     0.2700058            1.0000000            1.0000000
D   1   1.00
     0.8000000            1.0000000
****
-F     0
S   6   1.00
  7001.7130900            0.0018196169
  1051.3660900            0.0139160796
   239.2856900            0.0684053245
    67.3974453            0.233185760
    21.5199573            0.471267439
     7.40310130           0.356618546
SP   3   1.00
    20.8479528           -0.108506975          0.0716287243
     4.80830834          -0.146451658          0.3459121030
     1.34406986           1.128688580          0.7224699570
SP   1   1.00
     0.358151393          1.0000000            1.0000000
D   1   1.00
     0.8000000            1.0000000
****
-Si     0
S   6   1.00
 16115.9000000            0.00195948
  2425.5800000            0.01492880
   553.8670000            0.07284780
   156.3400000            0.24613000
    50.0683000            0.48591400
    17.0178000            0.32500200
SP   6   1.00
   292.7180000           -0.00278094           0.00443826
    69.8731000           -0.03571460           0.03266790
    22.3363000           -0.11498500           0.13472100
     8.1503900            0.09356340           0.32867800
     3.1345800            0.60301700           0.44964000
     1.2254300            0.41895900           0.26137200
SP   3   1.00
     1.7273800           -0.24463000          -0.01779510
     0.5729220            0.00431572           0.25353900
     0.2221920            1.09818000           0.80066900
SP   1   1.00
     0.0778369            1.00000000           1.00000000
D   1   1.00
     0.4500000            1.0000000
```

```
****
-P     0
S   6   1.00
  19413.3000000             0.0018516
   2909.4200000             0.0142062
    661.3640000             0.0699995
    185.7590000             0.2400790
     59.1943000             0.4847620
     20.0310000             0.3352000
SP   6   1.00
    339.4780000            -0.00278217            0.00456462
     81.0101000            -0.0360499             0.03369360
     25.8780000            -0.1166310             0.13975500
      9.4522100             0.0968328             0.33936200
      3.6656600             0.6144180             0.45092100
      1.4674600             0.4037980             0.23858600
SP   3   1.00
      2.1562300            -0.2529230            -0.01776530
      0.7489970             0.0328517             0.27405800
      0.2831450             1.0812500             0.78542100
SP   1   1.00
      0.0998317             1.0000000             1.00000000
D   1   1.00
      0.5500000             1.0000000
****
-S     0
S   6   1.00
  21917.1000000             0.0018690
   3301.4900000             0.0142300
    754.1460000             0.0696960
    212.7110000             0.2384870
     67.9896000             0.4833070
     23.0515000             0.3380740
SP   6   1.00
    423.7350000            -0.0023767             0.0040610
    100.7100000            -0.0316930             0.0306810
     32.1599000            -0.1133170             0.1304520
     11.8079000             0.0560900             0.3272050
      4.6311000             0.5922550             0.4528510
      1.8702500             0.4550060             0.2560420
SP   3   1.00
      2.6158400            -0.2503740            -0.0145110
      0.9221670             0.0669570             0.3102630
      0.3412870             1.0545100             0.7544830
SP   1   1.00
      0.1171670             1.0000000             1.0000000
D   1   1.00
      0.6500000             1.0000000
****
-Cl     0
S   6   1.00
  25180.1000000             0.0018330
   3780.3500000             0.0140340
    860.4740000             0.0690970
    242.1450000             0.2374520
     77.3349000             0.4830340
     26.2470000             0.3398560
SP   6   1.00
    491.7650000            -0.0022974             0.0039894
    116.9840000            -0.0307140             0.0303180
     37.4153000            -0.1125280             0.1298800
     13.7834000             0.0450160             0.3279510
      5.4521500             0.5893530             0.4535270
      2.2258800             0.4652060             0.2521540
```

```
SP   3   1.00
      3.1864900            -0.2518300            -0.0142990
      1.1442700             0.0615890             0.3235720
      0.4203770             1.0601800             0.7435070
SP   1   1.00
      0.1426570             1.0000000             1.0000000
D   1   1.00
      0.7500000             1.0000000
****
```

Section S5. ACP file for HF/MINIX

```
-H  0
H  1 0
local
8
2 0.120000 -0.028136546102616
2 0.140000 0.129052046084892
2 0.160000 -0.162834666416820
2 0.220000 0.115440203460067
2 0.300000 -0.051357765845977
2 0.500000 0.017033294338594
2 1.100000 -0.041676072842480
2 3.000000 -1.297558248426192
s
4
2 0.140000 0.018352544676934
2 0.220000 -0.135446994028465
2 0.700000 0.325211527432720
2 1.400000 0.415235236303971
-B  0
B  2 0
local
4
2 0.120000 -0.010431867828695
2 0.140000 0.022033004115136
2 0.260000 -0.006114693447935
2 0.700000 -0.043294428886973
s
1
2 3.000000 0.330524721538951
p
1
2 0.120000 -0.029603654772011
-C  0
C  2 0
local
7
2 0.120000 -0.035382604235113
2 0.140000 0.079325681924874
2 0.200000 -0.107554741449593
2 0.280000 0.092326283103470
2 0.400000 -0.065968911998717
2 0.600000 -0.226003613939323
2 0.900000 0.387374694506240
s
2
2 0.140000 0.056044730686507
2 0.220000 -0.246122437318296
p
5
2 0.120000 -0.050668778522707
2 0.180000 0.158957085517373
```

```
2 0.280000 0.148622386472062
2 0.800000 -0.147655448041361
2 1.300000 -0.575784044736115
-N  0
N  2 0
local
6
2 0.120000 -0.017517279343790
2 0.140000 0.053350090502697
2 0.180000 -0.068747457546253
2 0.300000 0.093048501457520
2 0.600000 -0.174690011868487
2 0.700000 -0.233973502650463
s
3
2 0.120000 -0.364185673569464
2 0.280000 -0.035059556580575
2 2.500000 1.958802809196974
p
4
2 0.160000 0.197827158897703
2 0.200000 0.160364082328380
2 0.700000 -0.346970871597540
2 3.000000 0.379794282290674
-O  0
O  2 0
local
7
2 0.120000 -0.030157675795425
2 0.140000 0.135303018500395
2 0.160000 -0.162581048481129
2 0.240000 0.130274411002281
2 0.400000 -0.089731000400052
2 0.600000 -0.133265805169002
2 2.500000 0.039849813462334
s
4
2 0.120000 0.259467960713428
2 0.180000 -0.312109139881752
2 0.500000 -0.150175452344375
2 1.700000 0.037267382091933
p
2
2 0.120000 -0.056805740131534
2 0.200000 0.076700585293407
-F  0
F  2 0
local
6
2 0.140000 -0.006826650517995
2 0.160000 -0.001840392648870
2 0.260000 0.036044341571709
2 0.280000 0.006128743221826
2 0.600000 -0.219544585232961
2 1.000000 -0.391060596224052
s
1
2 0.160000 -0.308078969617831
p
3
2 0.120000 0.087647805782508
2 0.280000 0.132237028553474
2 0.500000 0.055200629164079
-Si 0
```

```
Si 3 0
local
3
2 0.120000 -0.021461019780989
2 0.140000 0.051400342624278
2 0.220000 -0.024806786772559
s
1
2 0.280000 -0.576256241448157
p
1
2 0.120000 0.020223577126489
d
2
2 0.120000 -0.064941374697649
2 1.800000 -0.000000833235224
-P  0
P  3 0
local
4
2 0.120000 -0.034645639501834
2 0.140000 0.016504404247330
2 0.160000 0.023719143981247
2 0.600000 0.123897958778170
s
2
2 0.180000 -0.155669995312142
2 0.220000 -0.216144934783911
p
2
2 0.120000 0.020364118157026
2 0.220000 -0.244290574434349
d
4
2 0.120000 0.159724775487869
2 0.300000 -0.249770603584637
2 0.400000 -0.293067237363789
2 0.900000 -0.000017307708339
-S  0
S  3 0
local
5
2 0.120000 -0.042794818043893
2 0.140000 0.125943523362045
2 0.200000 -0.225505734887372
2 0.300000 0.260983352375191
2 0.900000 -0.357461158537498
s
3
2 0.120000 -0.027490230776148
2 0.300000 -0.000000439366168
2 0.500000 -0.000013218250161
p
1
2 0.120000 -0.022073069213358
d
1
2 0.120000 -0.090800450093055
-Cl 0
Cl 3 0
local
7
2 0.120000 -0.048326578687329
2 0.140000 0.084316305250296
```

```
2 0.160000 -0.000090475796118
2 0.240000 -0.010822013276623
2 0.400000 -0.216136710586582
2 0.500000 -0.000023127425500
2 1.400000 0.000006722842003
s
1
2 0.200000 -0.427322141292772
p
3
2 0.120000 0.035348354464651
2 0.220000 -0.000000871740192
2 0.500000 0.000021857099201
d
2
2 0.120000 0.010855487763334
2 0.140000 0.012772846052466
```

Section S6. ACP file for HF/6-31G*

```
-H  0
H  1 0
l
8
2 0.120000 -0.011979646752312
2 0.140000 0.044425137267345
2 0.180000 -0.090769495556353
2 0.240000 0.112659354385917
2 0.400000 -0.161566870491466
2 0.500000 -0.003357823290091
2 0.700000 0.258023863202355
2 2.500000 -1.114382209633294
s
5
2 0.120000 -0.079619001737395
2 0.160000 0.142632695063809
2 0.260000 -0.112654416806894
2 0.700000 0.272978591233067
2 1.400000 0.121256351999031
-B  0
B  3 0
l
3
2 0.120000 -0.001574337277511
2 0.160000 -0.001165192068335
2 0.500000 0.011026696097473
s
1
2 0.120000 0.016278024674017
p
1
2 0.300000 0.030966087642420
d
1
2 0.120000 -0.023015742409168
-C  0
C  3 0
l
7
2 0.120000 -0.019109049588823
2 0.140000 0.047043967807520
2 0.200000 -0.099179293154608
2 0.280000 0.121555604759121
```

```
2 0.400000 0.056212918757243
2 0.600000 -0.425902901640687
2 1.000000 0.836959381113923
s
3
2 0.120000 -0.018942735189471
2 0.200000 -0.072216690572218
2 0.500000 0.011017192265701
p
4
2 0.120000 -0.026488363615635
2 0.160000 0.023329358411437
2 0.300000 0.101063809927342
2 1.200000 -0.716266541937955
d
4
2 0.120000 -0.089227306042705
2 0.160000 0.319483984873804
2 0.220000 -0.349202991802217
2 0.400000 0.108277436222175
-N  0
N  3 0
l
6
2 0.120000 -0.038693653627767
2 0.140000 0.087412495270558
2 0.180000 -0.082345221017501
2 0.280000 0.101497367519688
2 0.500000 -0.009800198307370
2 0.600000 -0.283487242090359
s
2
2 0.140000 0.064954831847388
2 0.500000 0.160082154426835
p
3
2 0.120000 -0.123665716302443
2 0.220000 0.145607345141200
2 0.260000 0.161228580202816
d
4
2 0.120000 0.123981571252951
2 0.140000 0.028836546184427
2 0.220000 -0.550248697936328
2 0.700000 1.785419515176567
-O  0
O  3 0
l
7
2 0.120000 -0.024805536569800
2 0.140000 0.074386756754185
2 0.180000 -0.131017708654095
2 0.220000 0.049110494498567
2 0.260000 0.128980378002237
2 0.400000 -0.130162463439902
2 2.500000 -0.038456477761256
s
2
2 0.120000 0.029480290274379
2 1.900000 -0.587102105686877
p
3
2 0.120000 -0.011984192343963
2 0.160000 0.036734821706109
```

```
2 0.600000 -0.006681959105060
d
4
2 0.140000 0.075267012399339
2 0.220000 -0.147878274922841
2 0.240000 -0.187790720183225
2 0.800000 1.460490704611638
-F  0
F  3 0
l
5
2 0.120000 -0.004836482873288
2 0.140000 0.010021359561859
2 0.180000 -0.010527898272984
2 0.300000 0.023383178234878
2 0.800000 -0.098334724188930
s
3
2 0.120000 -0.026289336368737
2 0.240000 -0.341014550598012
2 0.280000 -0.049585222116591
p
2
2 0.120000 0.047055657129598
2 0.400000 -0.027192115248476
d
1
2 0.160000 0.045761183798386
-Si 0
Si 3 0
l
4
2 0.120000 -0.071704309528107
2 0.140000 0.146756051445767
2 0.200000 -0.132482954794748
2 0.400000 0.038037250517978
s
0
p
1
2 0.180000 0.053870455063610
d
0
-P  0
P  3 0
l
3
2 0.120000 -0.075551294878931
2 0.140000 0.103353491098827
2 0.220000 0.000560283408651
s
2
2 0.120000 0.049161488474092
2 0.240000 -0.231774434470701
p
1
2 0.160000 -0.078504243493890
d
3
2 0.120000 0.163491773328853
2 0.180000 -0.168269195236911
2 0.240000 -0.192540938977707
-S  0
S  3 0
```

```
l
4
2 0.120000 -0.031330854211208
2 0.140000 0.085205529589191
2 0.180000 -0.091936989104822
2 0.280000 0.073587353086563
s
2
2 0.120000 -0.162822221816308
2 0.400000 0.141040726565840
p
2
2 0.140000 0.009866786168893
2 0.800000 -0.170584340098399
d
2
2 0.120000 -0.051619430071470
2 0.200000 -0.013161146775379
-Cl 0
Cl 3 0
l
4
2 0.120000 -0.035142446802361
2 0.140000 0.068768418510911
2 0.180000 -0.015978662581669
2 0.800000 -0.030004751890848
s
2
2 0.120000 -0.037028050407337
2 0.180000 -0.316711816033988
p
1
2 0.220000 -0.026883211937695
d
1
2 0.200000 -0.036880059614231
```

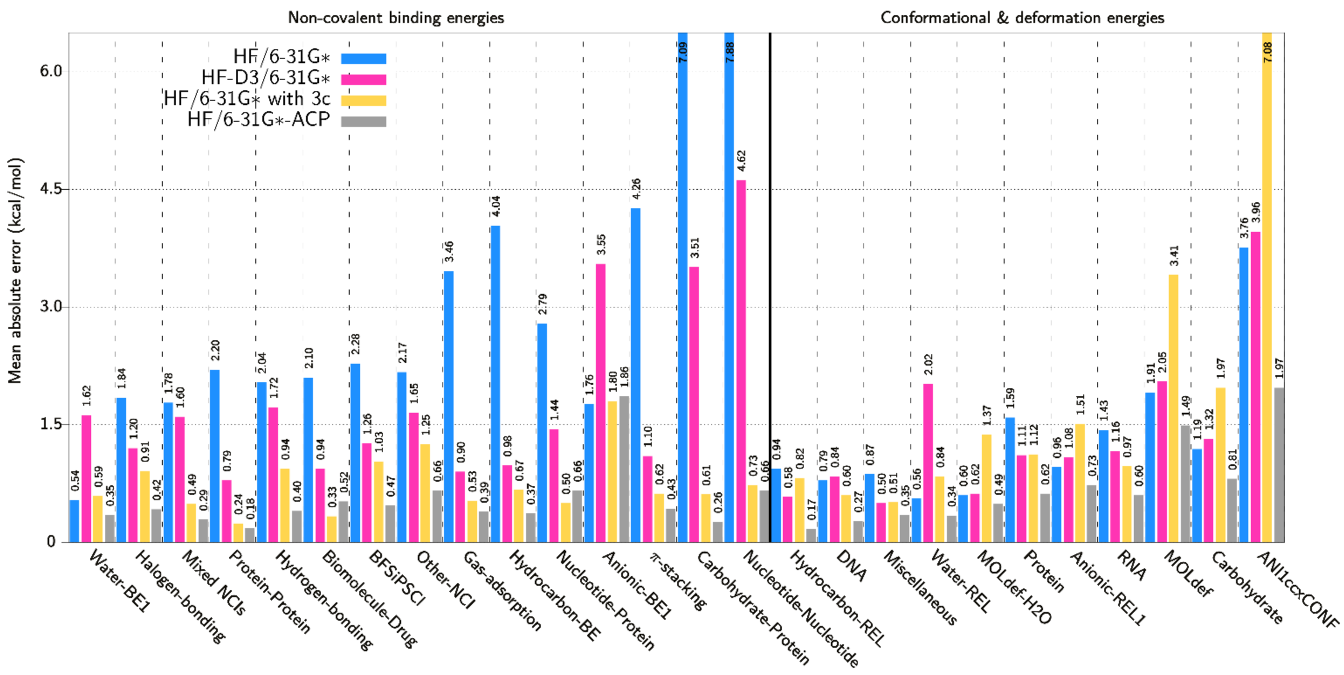


Figure S1. Mean absolute errors of HF/6-31G*-based methods (relative to the reference data) for the training set (Table 1). The methods shown are HF/6-31G* (blue), HF-D3/6-31G* (pink), HF/6-31G* with 3c (yellow), and HF/6-31G*-ACP (grey). The values for the mean absolute errors (in kcal/mol) are given atop the bars.

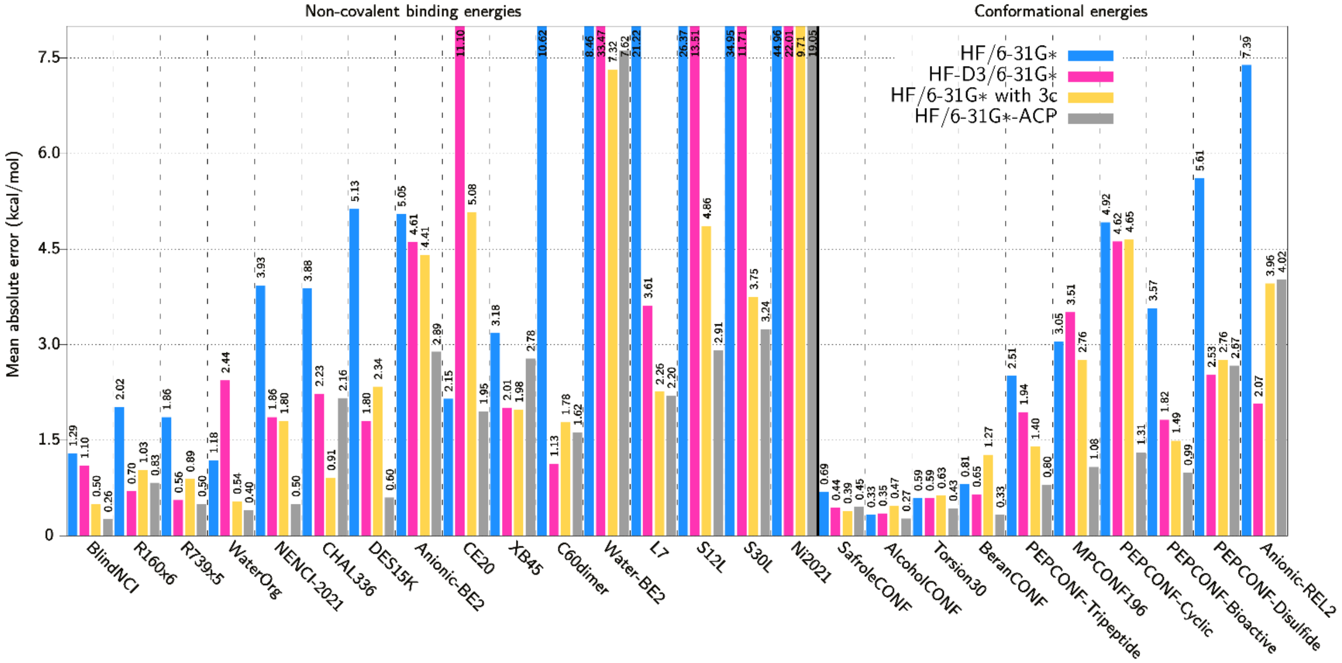


Figure S2. Mean absolute error of HF/6-31G*-based methods (relative to the reference data) for the validation set (Table 2). The methods shown are HF/6-31G* (blue), HF-D3/6-31G* (pink), HF/6-31G* with 3c (yellow), and HF/6-31G*-ACP (grey). The values for the mean absolute errors (in kcal/mol) are given atop the bars.

Table S1. The error analysis with respect to reference data of the various datasets present in the training set. The numbers in bracket in the first column indicates the number of datapoints. MAE = mean absolute error in kcal/mol, MSE = mean signed error in kcal/mol, MAXE = maximum absolute error in kcal/mol, RMSE = root-mean- square error in kcal/mol, SD = standard deviation in kcal/mol.

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
S22x5 (110)	MAE	2.41	1.40	0.53	0.46	2.22	1.56	0.70	0.34
	MSE	2.37	-1.18	0.17	0.27	2.00	-1.55	-0.20	0.00
	MAXE	18.89	8.26	2.62	1.52	20.28	5.40	3.04	1.21
	RMSE	3.95	2.42	0.80	0.59	4.16	2.21	0.99	0.44
	SD	3.18	2.12	0.79	0.53	3.66	1.57	0.98	0.44
S66x8 (528)	MAE	2.29	1.24	0.37	0.38	2.05	1.49	0.50	0.29
	MSE	2.25	-1.15	0.08	0.14	1.91	-1.49	-0.26	0.02
	MAXE	12.38	6.68	2.46	1.35	12.83	5.09	3.02	1.03
	RMSE	3.09	1.78	0.56	0.48	3.12	1.88	0.67	0.36
	SD	2.12	1.36	0.55	0.46	2.47	1.15	0.62	0.36

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
S66a8 (528)	MAE	2.25	1.18	0.36	0.33	1.84	1.64	0.44	0.26
	MSE	2.21	-1.18	0.06	0.11	1.75	-1.64	-0.40	0.00
	MAXE	6.82	3.93	4.07	2.91	5.36	5.19	1.63	1.19
	RMSE	2.55	1.37	0.54	0.47	2.24	1.91	0.55	0.33
	SD	1.27	0.70	0.54	0.45	1.39	0.98	0.37	0.33
A21x12 (228)	MAE	0.45	0.26	0.19	0.14	0.34	0.44	0.23	0.14
	MSE	0.39	-0.19	0.12	0.05	0.14	-0.44	-0.13	-0.04
	MAXE	4.15	3.84	3.27	1.71	4.63	3.08	2.31	1.08
	RMSE	0.82	0.62	0.42	0.25	0.69	0.78	0.41	0.22
	SD	0.73	0.59	0.40	0.25	0.67	0.65	0.39	0.22
NBC10ext (195)	MAE	3.03	1.25	0.54	0.44	3.69	0.60	0.40	0.33
	MSE	3.03	-1.24	-0.50	-0.08	3.69	-0.58	0.16	-0.14
	MAXE	8.00	5.05	3.30	1.95	11.96	2.40	3.64	1.05
	RMSE	3.69	1.80	0.90	0.61	4.85	0.83	0.64	0.42
	SD	2.12	1.30	0.75	0.61	3.16	0.59	0.62	0.40
Sulfurx8 (104)	MAE	1.75	0.36	0.71	0.35	1.27	0.75	0.40	0.23
	MSE	1.75	-0.24	0.66	0.19	1.25	-0.74	0.16	0.07
	MAXE	6.54	2.01	5.13	1.69	6.01	2.31	2.86	0.87
	RMSE	2.45	0.57	1.28	0.44	1.90	0.96	0.65	0.30
	SD	1.72	0.52	1.10	0.40	1.43	0.62	0.63	0.29
3B-69-DIM (207)	MAE	1.85	1.08	0.50	0.45	1.42	1.44	0.46	0.27
	MSE	1.80	-1.01	0.28	0.27	1.38	-1.43	-0.14	0.13
	MAXE	5.59	6.12	3.18	4.51	6.34	4.41	1.92	1.33
	RMSE	2.37	1.68	0.72	0.70	1.93	1.80	0.60	0.37
	SD	1.55	1.34	0.67	0.65	1.36	1.10	0.59	0.34
3B-69-TRIM (69)	MAE	5.45	3.14	1.13	1.02	4.22	4.32	1.11	0.60
	MSE	5.30	-3.12	0.75	0.74	4.10	-4.32	-0.44	0.39
	MAXE	11.31	9.00	4.96	7.63	12.86	10.07	3.33	2.21
	RMSE	6.17	3.93	1.50	1.50	4.99	4.69	1.33	0.83
	SD	3.19	2.40	1.31	1.31	2.87	1.84	1.27	0.74
WatAA (259)	MAE	1.37	3.54	1.25	0.81	0.63	3.29	0.65	0.35
	MSE	-0.03	-3.54	-0.79	-0.18	0.23	-3.29	-0.54	-0.21
	MAXE	4.95	8.08	5.35	3.15	2.64	7.52	2.46	1.36
	RMSE	1.82	3.90	1.80	1.10	0.78	3.41	0.82	0.44
	SD	1.83	1.63	1.62	1.09	0.75	0.91	0.62	0.38
BBI (100)	MAE	2.56	1.04	0.88	1.27	1.91	1.69	0.45	0.62
	MSE	2.56	-1.03	0.88	1.27	1.91	-1.67	0.24	0.62
	MAXE	4.12	2.33	1.79	2.71	3.99	2.64	1.46	1.01
	RMSE	2.65	1.18	0.92	1.39	2.17	1.79	0.55	0.64
	SD	0.69	0.57	0.25	0.57	1.02	0.64	0.50	0.15
SSI (2805)	MAE	2.14	0.87	0.28	0.26	2.21	0.76	0.24	0.16
	MSE	2.13	-0.82	-0.03	0.18	2.20	-0.75	0.04	0.04
	MAXE	10.76	6.86	4.76	3.67	14.04	4.45	7.65	2.71
	RMSE	2.44	1.05	0.49	0.42	2.47	0.95	0.40	0.24

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
	SD	1.20	0.66	0.49	0.38	1.12	0.58	0.40	0.24
JSCH (124)	MAE	5.07	2.56	0.98	0.70	5.25	2.30	0.70	0.49
	MSE	5.00	-2.48	0.15	0.05	5.25	-2.24	0.39	0.21
	MAXE	17.58	13.82	6.12	3.82	17.13	9.05	2.99	1.80
	RMSE	6.36	3.76	1.36	0.96	6.32	3.13	0.93	0.61
	SD	3.95	2.84	1.36	0.96	3.54	2.19	0.85	0.57
DNAstack (40)	MAE	4.67	0.87	0.42	0.52	4.66	0.90	0.43	0.43
	MSE	4.67	-0.77	0.37	0.32	4.66	-0.78	0.36	0.20
	MAXE	9.91	2.42	1.33	1.86	10.73	2.19	2.06	1.16
	RMSE	5.64	1.12	0.55	0.65	5.68	1.12	0.60	0.48
	SD	3.19	0.82	0.40	0.57	3.28	0.81	0.49	0.44
DNA2body (10)	MAE	17.27	4.48	0.21	0.37	18.12	3.63	0.95	0.40
	MSE	17.27	-4.48	0.09	-0.08	18.12	-3.63	0.95	0.24
	MAXE	18.15	5.16	0.37	0.73	19.09	4.48	1.66	0.89
	RMSE	17.28	4.49	0.24	0.42	18.13	3.66	1.00	0.49
	SD	0.58	0.30	0.23	0.44	0.74	0.48	0.33	0.45
ACHC (54)	MAE	6.89	1.44	0.28	0.57	7.18	1.15	0.36	0.48
	MSE	6.89	-1.44	0.06	-0.39	7.18	-1.15	0.35	-0.35
	MAXE	11.83	5.62	1.48	1.48	15.09	2.36	1.78	1.07
	RMSE	7.05	1.66	0.37	0.68	7.41	1.22	0.47	0.54
	SD	1.49	0.83	0.36	0.55	1.84	0.43	0.32	0.42
BDNA (71)	MAE	4.08	2.08	0.60	0.55	4.48	1.69	0.38	0.32
	MSE	4.08	-2.05	-0.09	0.00	4.48	-1.65	0.31	0.18
	MAXE	10.77	7.81	2.22	1.10	12.38	4.57	1.72	0.75
	RMSE	5.14	3.11	0.81	0.64	5.49	2.30	0.55	0.37
	SD	3.15	2.36	0.81	0.64	3.20	1.62	0.45	0.33
NucBTrimer (141)	MAE	12.73	10.27	1.73	1.54	12.36	10.60	1.14	1.14
	MSE	12.69	-10.27	0.17	0.84	12.36	-10.60	-0.16	0.76
	MAXE	33.69	19.17	6.79	9.07	32.57	17.12	4.38	3.37
	RMSE	15.15	10.71	2.35	2.01	14.31	10.87	1.43	1.32
	SD	8.31	3.05	2.35	1.84	7.23	2.42	1.43	1.08
NucTAA (454)	MAE	3.50	1.22	0.92	1.15	2.79	1.44	0.50	0.66
	MSE	3.39	-0.75	0.62	1.01	2.77	-1.38	-0.01	0.52
	MAXE	13.86	18.78	12.56	7.32	14.13	7.49	5.46	4.39
	RMSE	4.15	2.13	1.54	1.50	3.56	1.89	0.77	0.86
	SD	2.39	1.99	1.42	1.12	2.24	1.30	0.77	0.68
CarbhydBz (34)	MAE	6.85	2.96	0.62	0.48	7.27	2.53	0.25	0.18
	MSE	6.85	-2.96	-0.54	-0.34	7.27	-2.53	-0.12	0.00
	MAXE	7.57	3.69	0.95	1.16	8.31	3.36	1.18	1.13
	RMSE	6.86	3.01	0.67	0.53	7.29	2.58	0.32	0.26
	SD	0.42	0.54	0.39	0.41	0.47	0.49	0.31	0.26
CarbhydNaph (46)	MAE	9.39	3.72	0.61	0.40	10.13	2.97	0.23	0.37
	MSE	9.39	-3.72	-0.61	-0.37	10.13	-2.97	0.13	0.37
	MAXE	12.07	4.83	0.94	0.88	12.90	4.26	0.73	0.81

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
	RMSE	9.54	3.77	0.64	0.46	10.27	3.05	0.29	0.41
	SD	1.71	0.66	0.19	0.28	1.73	0.69	0.26	0.19
CarbhydAroAA (48)	MAE	4.88	1.61	0.35	0.56	3.94	2.55	1.19	0.36
	MSE	4.88	-1.61	-0.25	0.49	3.94	-2.55	-1.19	-0.04
	MAXE	8.53	3.64	0.69	1.46	7.65	5.15	2.45	1.09
	RMSE	5.26	1.82	0.40	0.64	4.30	2.78	1.31	0.43
	SD	1.98	0.86	0.31	0.41	1.75	1.12	0.55	0.43
CarbhydAro (161)	MAE	7.23	4.05	0.56	0.42	7.12	4.16	0.63	0.22
	MSE	7.23	-4.05	-0.52	-0.34	7.12	-4.16	-0.62	-0.04
	MAXE	9.28	7.25	1.34	1.33	9.10	7.42	1.61	0.83
	RMSE	7.34	4.23	0.64	0.52	7.23	4.35	0.70	0.27
	SD	1.27	1.22	0.39	0.40	1.25	1.28	0.33	0.27
HSG (17)	MAE	2.51	0.94	0.33	0.28	2.48	0.89	0.38	0.24
	MSE	2.51	-0.86	0.08	0.23	2.48	-0.89	0.05	0.15
	MAXE	3.57	1.72	1.36	0.83	4.35	2.64	0.81	0.60
	RMSE	2.58	1.01	0.47	0.37	2.66	1.04	0.44	0.30
	SD	0.63	0.56	0.48	0.30	1.00	0.56	0.46	0.27
PLF547 (392)	MAE	2.39	0.78	0.47	0.75	2.09	0.94	0.33	0.53
	MSE	2.39	-0.56	0.30	0.71	2.08	-0.87	-0.01	0.50
	MAXE	15.22	6.14	5.91	4.69	10.91	6.54	2.85	2.66
	RMSE	3.30	1.27	0.82	1.00	2.93	1.40	0.52	0.68
	SD	2.28	1.14	0.77	0.72	2.07	1.10	0.52	0.45
HBC6 (118)	MAE	1.14	3.27	1.13	0.58	1.47	2.79	1.27	0.41
	MSE	0.86	-2.84	-0.01	0.10	0.91	-2.79	0.04	-0.19
	MAXE	3.49	11.34	4.75	1.91	7.97	5.43	4.71	1.56
	RMSE	1.45	4.65	1.48	0.75	2.32	3.33	1.65	0.53
	SD	1.18	3.70	1.48	0.74	2.14	1.82	1.66	0.49
MiriyalaHB104 (104)	MAE	1.49	2.41	0.52	0.49	1.40	2.44	0.43	0.22
	MSE	1.33	-2.41	-0.28	-0.24	1.29	-2.44	-0.32	0.01
	MAXE	4.18	5.63	2.01	2.06	4.07	5.15	1.52	1.08
	RMSE	1.69	2.59	0.67	0.62	1.66	2.58	0.51	0.29
	SD	1.06	0.95	0.62	0.58	1.05	0.83	0.40	0.29
IonicHB (96)	MAE	2.32	4.50	2.72	1.85	1.07	2.94	1.13	0.46
	MSE	-1.68	-4.34	-2.38	-1.70	-0.28	-2.94	-0.98	-0.42
	MAXE	8.46	14.62	10.73	7.34	3.55	5.28	2.57	1.19
	RMSE	3.21	5.82	3.79	2.39	1.27	3.13	1.29	0.55
	SD	2.75	3.91	2.97	1.68	1.25	1.07	0.85	0.36
HB375x10 (3749)	MAE	2.15	1.68	0.53	0.54	1.99	1.81	0.67	0.30
	MSE	2.11	-1.62	0.18	0.14	1.92	-1.81	-0.01	0.03
	MAXE	13.15	7.49	5.85	8.41	17.57	6.11	5.24	2.57
	RMSE	2.80	2.29	0.77	0.81	2.90	2.11	0.91	0.41
	SD	1.84	1.62	0.75	0.80	2.17	1.08	0.91	0.41
IHB100x10 (350)	MAE	2.37	4.72	2.72	2.19	2.46	2.25	1.41	0.77
	MSE	-0.06	-4.19	-1.71	-0.74	2.15	-1.98	0.50	-0.29

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
	MAXE	9.47	17.17	11.84	8.31	10.24	4.79	7.88	3.40
	RMSE	3.08	6.18	3.77	2.85	3.28	2.59	1.83	1.01
	SD	3.08	4.56	3.37	2.76	2.48	1.66	1.76	0.97
HB300SPXx10 (1980)	MAE	2.82	1.44	1.98	0.83	2.17	1.30	1.36	0.52
	MSE	2.59	-0.69	1.69	0.02	2.02	-1.26	1.13	0.08
	MAXE	18.20	19.39	21.15	15.28	12.86	7.67	10.44	3.92
	RMSE	3.97	2.50	3.35	1.48	3.12	1.79	2.19	0.73
	SD	3.02	2.41	2.90	1.48	2.37	1.28	1.88	0.73
Pisub (105)	MAE	7.22	2.02	0.57	0.42	8.22	1.02	0.71	0.54
	MSE	7.22	-2.02	-0.36	0.10	8.22	-1.02	0.64	-0.48
	MAXE	14.18	4.77	1.62	2.21	15.70	3.26	2.80	1.31
	RMSE	7.82	2.17	0.69	0.59	8.86	1.20	0.90	0.63
	SD	3.01	0.81	0.60	0.59	3.31	0.64	0.63	0.41
Pi29n (29)	MAE	5.92	0.84	0.29	0.71	6.40	0.44	0.59	0.42
	MSE	5.92	-0.84	-0.01	0.67	6.40	-0.36	0.48	0.27
	MAXE	23.72	2.86	1.44	1.99	22.39	4.19	1.29	1.50
	RMSE	7.31	1.02	0.38	0.86	7.72	0.91	0.68	0.54
	SD	4.37	0.59	0.39	0.55	4.39	0.86	0.49	0.48
BzDC215 (170)	MAE	1.89	0.85	0.23	0.50	1.45	1.27	0.58	0.36
	MSE	1.89	-0.82	-0.03	-0.09	1.44	-1.27	-0.48	-0.18
	MAXE	7.15	4.22	1.45	2.13	7.49	4.79	1.67	1.26
	RMSE	2.43	1.30	0.36	0.68	2.21	1.70	0.77	0.47
	SD	1.53	1.02	0.36	0.67	1.68	1.13	0.60	0.44
Hill18 (18)	MAE	3.91	1.90	2.30	0.56	3.17	1.92	1.96	0.65
	MSE	3.91	-0.31	2.30	0.16	3.17	-1.05	1.56	0.07
	MAXE	21.80	12.90	17.58	1.55	16.66	7.77	12.45	2.59
	RMSE	6.11	3.36	4.54	0.68	4.90	2.45	3.38	0.90
	SD	4.83	3.44	4.03	0.68	3.84	2.28	3.09	0.93
X40x10 (220)	MAE	1.75	1.41	0.86	0.63	1.74	1.14	0.82	0.40
	MSE	1.29	-1.35	0.37	-0.22	1.50	-1.14	0.58	0.01
	MAXE	17.76	8.13	7.43	5.16	23.39	4.31	7.69	2.25
	RMSE	3.05	2.42	1.39	1.11	3.49	1.60	1.36	0.56
	SD	2.77	2.01	1.34	1.09	3.16	1.13	1.23	0.56
PNICO23 (23)	MAE	2.84	2.17	1.80	0.78	2.77	1.64	1.44	0.62
	MSE	-2.18	1.74	0.28	0.49	-2.65	1.28	-0.19	0.34
	MAXE	15.00	9.02	10.08	3.18	14.20	5.34	9.28	1.88
	RMSE	4.16	3.22	3.01	1.12	4.11	2.29	2.39	0.78
	SD	3.62	2.77	3.07	1.03	3.21	1.95	2.44	0.72
CARBHB12 (12)	MAE	1.62	1.10	0.68	1.02	1.03	1.66	0.88	0.74
	MSE	-1.48	1.10	-0.09	-0.31	-0.92	1.66	0.48	0.25
	MAXE	3.57	3.69	2.10	2.86	3.30	3.38	1.70	1.48
	RMSE	1.87	1.59	0.92	1.30	1.47	1.91	1.05	0.87
	SD	1.19	1.20	0.96	1.32	1.20	0.99	0.97	0.86
ADIM6 (6)	MAE	4.54	1.90	0.47	0.13	5.29	1.15	0.28	0.18

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
	MSE	4.54	-1.90	-0.47	0.13	5.29	-1.15	0.28	0.14
	MAXE	7.77	3.16	0.71	0.39	9.07	1.86	0.63	0.57
	RMSE	4.98	2.09	0.52	0.18	5.82	1.24	0.34	0.27
	SD	2.25	0.95	0.24	0.13	2.67	0.52	0.21	0.25
HC12 (12)	MAE	4.56	1.80	0.42	0.15	5.17	1.19	0.51	0.34
	MSE	4.56	-1.80	-0.42	0.06	5.17	-1.19	0.19	-0.17
	MAXE	7.30	3.39	0.91	0.58	8.86	2.54	1.31	0.71
	RMSE	4.73	1.96	0.51	0.21	5.45	1.30	0.59	0.42
	SD	1.33	0.81	0.30	0.21	1.81	0.55	0.58	0.41
HW30 (30)	MAE	1.10	0.81	0.31	0.32	0.51	1.49	0.66	0.38
	MSE	1.07	-0.80	0.05	-0.19	0.37	-1.49	-0.65	-0.22
	MAXE	2.69	1.86	0.82	0.98	1.31	3.63	2.08	1.29
	RMSE	1.38	0.95	0.39	0.41	0.62	1.73	0.89	0.52
	SD	0.88	0.53	0.39	0.37	0.50	0.89	0.63	0.48
C2H4NT (75)	MAE	4.20	1.71	0.48	0.72	5.17	0.73	0.73	0.38
	MSE	4.20	-1.70	-0.45	0.37	5.17	-0.73	0.52	0.34
	MAXE	12.76	5.52	2.06	2.05	16.15	1.62	3.75	0.97
	RMSE	5.14	2.53	0.70	0.82	6.76	0.86	1.23	0.46
	SD	2.98	1.89	0.54	0.74	4.39	0.46	1.12	0.30
CH4PAH (382)	MAE	3.09	1.19	0.19	0.32	3.54	0.74	0.43	0.27
	MSE	3.09	-1.19	-0.19	0.14	3.54	-0.74	0.26	0.04
	MAXE	13.50	6.94	1.58	0.99	16.27	4.17	2.92	1.62
	RMSE	4.01	1.84	0.32	0.38	4.83	1.00	0.76	0.35
	SD	2.56	1.40	0.26	0.35	3.29	0.68	0.71	0.35
CO2MOF (20)	MAE	1.76	2.18	0.92	0.84	1.37	2.55	1.16	0.83
	MSE	1.74	-2.14	-0.71	-0.74	1.33	-2.55	-1.12	-0.67
	MAXE	4.49	4.45	2.41	2.28	3.71	4.29	2.86	1.75
	RMSE	2.19	2.50	1.27	1.12	1.79	2.79	1.45	1.00
	SD	1.36	1.32	1.08	0.87	1.23	1.16	0.93	0.77
CO2PAH (249)	MAE	3.82	1.64	0.55	0.68	4.51	0.97	0.68	0.58
	MSE	3.82	-1.59	-0.32	0.58	4.51	-0.90	0.37	0.54
	MAXE	13.08	9.01	2.70	2.80	16.68	5.44	3.52	1.89
	RMSE	5.03	2.62	0.82	0.92	6.25	1.47	1.04	0.73
	SD	3.28	2.09	0.76	0.71	4.33	1.16	0.97	0.49
CO2NPHAC (96)	MAE	2.10	1.92	0.63	0.44	2.64	1.38	0.63	0.41
	MSE	2.10	-1.92	-0.58	0.24	2.64	-1.38	-0.04	0.30
	MAXE	10.35	11.42	6.59	2.72	14.94	4.10	3.99	2.42
	RMSE	3.01	3.04	1.19	0.69	4.18	1.81	0.95	0.62
	SD	2.16	2.37	1.05	0.65	3.26	1.17	0.95	0.55
BzGas (129)	MAE	1.81	0.92	0.35	0.45	2.12	0.63	0.37	0.31
	MSE	1.81	-0.88	-0.21	0.02	2.12	-0.57	0.10	0.00
	MAXE	4.49	6.13	3.01	3.13	9.17	1.54	2.15	1.24
	RMSE	2.05	1.41	0.60	0.62	2.69	0.73	0.50	0.39
	SD	0.95	1.11	0.56	0.63	1.66	0.45	0.49	0.39

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
Water38 (38)	MAE	12.76	30.62	7.67	1.67	1.66	19.51	3.43	0.49
	MSE	-12.76	-30.62	-7.67	-1.57	-1.66	-19.51	3.43	0.00
	MAXE	24.89	60.84	15.61	3.23	2.89	38.12	7.26	1.53
	RMSE	13.96	33.29	8.53	1.82	1.79	21.17	3.90	0.66
	SD	5.73	13.25	3.77	0.94	0.68	8.33	1.88	0.66
Water1888 (1888)	MAE	0.72	1.44	0.70	0.68	0.60	1.48	0.62	0.41
	MSE	0.04	-1.32	0.03	-0.04	0.01	-1.34	0.00	0.02
	MAXE	4.42	5.92	5.01	4.95	2.64	4.53	2.91	2.01
	RMSE	0.92	1.85	0.91	0.89	0.76	1.75	0.80	0.56
	SD	0.92	1.30	0.91	0.88	0.76	1.12	0.80	0.56
Water-2body (410)	MAE	0.23	0.63	0.17	0.13	0.17	0.65	0.19	0.09
	MSE	-0.02	-0.56	-0.02	0.05	-0.05	-0.59	-0.05	-0.02
	MAXE	1.11	3.59	0.85	0.67	0.98	3.27	1.11	0.82
	RMSE	0.37	1.26	0.25	0.20	0.25	1.12	0.27	0.15
	SD	0.37	1.13	0.25	0.19	0.24	0.96	0.26	0.14
B-set (160)	MAE	3.41	3.55	2.55	1.01	3.80	1.70	1.21	0.51
	MSE	1.96	-2.99	-1.07	-0.16	3.57	-1.37	0.55	-0.06
	MAXE	30.42	16.77	24.77	5.46	20.93	7.75	15.28	1.98
	RMSE	5.30	5.77	4.98	1.60	5.61	2.41	2.48	0.68
	SD	4.94	4.95	4.88	1.60	4.35	1.98	2.42	0.67
F-set (160)	MAE	1.84	1.37	1.37	0.57	1.40	1.08	1.50	0.37
	MSE	0.95	-1.11	1.22	0.21	1.15	-0.91	1.42	0.05
	MAXE	15.28	6.74	14.90	5.98	12.69	4.25	12.30	1.67
	RMSE	2.99	2.03	2.69	0.90	2.57	1.47	2.54	0.49
	SD	2.85	1.71	2.41	0.88	2.30	1.16	2.11	0.49
Si-set (152)	MAE	1.80	2.39	1.19	0.62	2.07	1.99	0.93	0.48
	MSE	1.38	-2.37	-0.91	-0.21	1.76	-1.99	-0.53	-0.09
	MAXE	10.46	13.71	8.68	3.79	13.30	9.74	5.79	3.96
	RMSE	2.75	4.13	2.30	0.93	3.47	3.06	1.52	0.79
	SD	2.39	3.39	2.12	0.91	3.00	2.33	1.43	0.79
P-set (120)	MAE	2.14	1.26	1.04	0.72	1.62	1.01	0.73	0.48
	MSE	1.99	-0.48	0.64	0.12	1.55	-0.91	0.20	0.17
	MAXE	9.16	15.53	6.58	3.49	12.01	5.48	7.11	1.87
	RMSE	2.93	2.56	1.63	1.07	2.66	1.51	1.42	0.66
	SD	2.16	2.52	1.51	1.07	2.17	1.20	1.41	0.64
S-set (144)	MAE	2.23	0.50	0.70	0.42	1.89	0.79	0.19	0.36
	MSE	2.21	-0.44	0.51	0.09	1.88	-0.77	0.48	0.13
	MAXE	12.55	4.36	6.57	2.88	14.24	2.55	6.03	1.79
	RMSE	3.24	0.87	1.30	0.60	3.04	1.03	0.88	0.51
	SD	2.38	0.76	1.20	0.59	2.40	0.69	0.86	0.50
Cl-set (160)	MAE	2.89	1.20	0.81	0.72	2.66	0.91	1.21	0.60
	MSE	1.93	-1.07	0.26	0.05	2.63	-0.37	0.96	0.42
	MAXE	26.97	9.01	7.82	5.85	27.87	5.40	13.46	4.52
	RMSE	4.97	2.31	1.33	1.11	4.84	1.44	2.72	0.99

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
	SD	4.60	2.06	1.31	1.11	4.08	1.40	2.55	0.90
SSI (anionic) (575)	MAE	3.42	3.68	2.83	2.17	1.64	3.06	1.47	1.55
	MSE	0.88	-2.45	-0.48	0.21	0.58	-2.75	-0.78	-0.98
	MAXE	10.75	18.25	12.06	8.97	8.66	9.18	7.74	7.52
	RMSE	3.89	5.18	3.56	2.70	2.11	3.82	1.90	2.08
	SD	3.80	4.57	3.53	2.70	2.03	2.66	1.74	1.83
WatAA (anionic) (64)	MAE	2.97	6.33	3.18	1.64	0.58	4.08	0.85	0.86
	MSE	-2.63	-6.33	-3.09	-1.51	-0.38	-4.08	-0.84	-0.83
	MAXE	5.05	8.75	5.12	3.74	1.45	6.97	2.47	1.74
	RMSE	3.30	6.59	3.51	1.85	0.72	4.15	0.97	0.95
	SD	2.01	1.84	1.66	1.07	0.62	0.78	0.49	0.47
HSG (anionic) (4)	MAE	2.49	4.88	2.47	2.08	0.31	4.42	1.13	0.95
	MSE	-0.07	-4.56	-1.27	-0.94	0.07	-4.42	-1.13	-0.90
	MAXE	3.23	10.29	4.23	4.01	0.47	6.84	1.82	1.48
	RMSE	2.55	6.45	2.80	2.39	0.32	4.92	1.22	1.09
	SD	2.94	5.27	2.88	2.54	0.36	2.50	0.54	0.70
PLF547 (anionic) (155)	MAE	3.09	2.57	1.78	1.55	1.85	2.09	0.93	1.04
	MSE	1.21	-2.14	-0.74	0.04	1.40	-1.95	-0.55	-0.20
	MAXE	15.61	23.84	16.83	10.86	10.78	11.58	6.11	6.76
	RMSE	4.23	5.41	3.71	2.57	2.61	3.45	1.56	1.69
	SD	4.07	4.99	3.64	2.58	2.21	2.86	1.46	1.69
IonicHB (anionic) (24)	MAE	2.06	4.23	2.21	1.80	1.73	4.38	2.18	1.69
	MSE	-1.38	-4.03	-1.82	-1.42	-1.73	-4.38	-2.18	-1.65
	MAXE	6.24	11.80	6.25	4.76	3.98	8.09	4.34	2.69
	RMSE	2.76	5.53	2.94	2.34	2.11	4.89	2.45	1.84
	SD	2.44	3.87	2.36	1.90	1.22	2.24	1.14	0.82
IHB100x10 (anionic) (650)	MAE	4.18	6.51	4.58	4.04	1.84	4.20	2.25	2.39
	MSE	-3.30	-6.19	-4.04	-3.55	-1.30	-4.20	-2.05	-2.36
	MAXE	32.63	35.49	32.36	26.47	12.98	16.20	14.12	11.26
	RMSE	6.73	9.20	7.12	6.45	2.74	4.95	3.05	3.17
	SD	5.87	6.81	5.87	5.38	2.42	2.63	2.26	2.12
Ionic43 (anionic) (37)	MAE	6.67	10.85	6.70	5.46	4.12	4.26	4.19	2.69
	MSE	-6.34	-10.85	-6.12	-5.28	0.54	-3.97	0.77	-2.28
	MAXE	44.77	46.30	42.06	30.90	17.19	18.73	14.49	13.08
	RMSE	12.05	14.55	11.51	8.72	5.65	6.05	5.50	4.34
	SD	10.38	9.82	9.88	7.03	5.70	4.63	5.52	3.74
PEPCONF-Dipeptide (875)	MAE	2.40	1.84	1.26	0.97	1.67	1.28	1.03	0.58
	MSE	-1.19	0.76	-0.05	-0.10	-0.86	1.09	0.27	0.18
	MAXE	10.23	11.39	7.60	4.83	7.82	5.32	3.82	2.77
	RMSE	3.10	2.41	1.78	1.25	2.25	1.62	1.28	0.73
	SD	2.87	2.29	1.78	1.25	2.08	1.20	1.25	0.71
TPCONF (8)	MAE	1.50	5.14	3.64	0.45	2.48	2.15	0.59	0.38
	MSE	-1.03	-5.14	-3.64	-0.32	2.48	-1.63	-0.13	-0.28

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
	MAXE	3.92	11.97	8.23	1.85	4.60	3.74	1.51	0.74
	RMSE	1.82	6.49	4.45	0.73	3.18	2.37	0.81	0.43
	SD	1.60	4.24	2.73	0.69	2.13	1.84	0.86	0.35
P76 (71)	MAE	2.70	2.07	1.24	1.06	2.34	1.04	2.28	0.59
	MSE	1.67	2.02	0.63	-0.07	-0.15	0.20	-1.20	-0.05
	MAXE	11.99	7.36	4.60	3.11	7.39	3.74	7.05	2.01
	RMSE	3.69	2.65	1.67	1.28	3.00	1.38	3.02	0.76
	SD	3.32	1.72	1.55	1.29	3.02	1.37	2.79	0.77
YMPJ (495)	MAE	2.09	1.74	2.32	1.15	1.32	0.80	1.14	0.69
	MSE	-1.85	-1.17	-2.27	-0.57	-0.12	0.57	-0.53	-0.59
	MAXE	6.31	6.29	6.90	4.22	5.67	2.89	4.35	2.26
	RMSE	2.60	2.15	2.70	1.43	1.76	1.01	1.43	0.83
	SD	1.83	1.81	1.48	1.31	1.76	0.83	1.33	0.58
SPS (17)	MAE	0.74	1.69	0.58	0.55	0.79	0.84	0.60	0.27
	MSE	0.54	1.67	0.49	0.33	-0.42	0.71	-0.48	0.12
	MAXE	2.19	4.77	2.63	1.87	1.46	1.69	1.95	0.49
	RMSE	0.95	2.05	0.91	0.70	0.94	0.93	0.81	0.30
	SD	0.80	1.22	0.79	0.64	0.87	0.62	0.68	0.29
rSPS (45)	MAE	1.36	1.02	1.14	0.69	0.76	0.65	0.73	0.34
	MSE	-1.24	-0.57	-0.92	-0.54	-0.38	0.29	-0.06	-0.18
	MAXE	2.84	3.14	2.84	1.99	2.60	2.60	1.93	1.27
	RMSE	1.55	1.29	1.33	0.85	1.03	0.83	0.85	0.45
	SD	0.95	1.17	0.97	0.66	0.96	0.78	0.86	0.42
UpU46 (45)	MAE	4.14	3.18	2.94	1.46	2.10	1.67	1.22	0.86
	MSE	4.08	3.06	2.84	1.34	1.77	0.75	0.53	0.45
	MAXE	11.30	6.76	6.12	5.32	6.63	5.16	4.01	3.60
	RMSE	5.14	3.69	3.35	1.83	2.94	2.16	1.59	1.12
	SD	3.15	2.09	1.80	1.27	2.37	2.05	1.52	1.03
SCONF (17)	MAE	2.63	5.20	1.47	0.68	1.05	1.57	2.19	0.78
	MSE	0.15	1.51	-0.42	-0.48	-0.60	0.76	-1.17	-0.69
	MAXE	9.81	14.76	7.63	3.88	2.22	3.43	5.02	1.38
	RMSE	3.72	6.35	2.57	1.17	1.20	1.76	2.46	0.87
	SD	3.83	6.35	2.61	1.10	1.07	1.64	2.23	0.55
DSCONF (27)	MAE	3.18	5.30	2.47	1.40	1.07	3.31	1.58	0.53
	MSE	-0.41	2.32	-0.08	0.65	0.42	3.15	0.75	0.18
	MAXE	6.56	13.69	5.24	3.22	2.66	6.62	4.69	1.40
	RMSE	3.70	6.27	2.88	1.69	1.34	3.65	1.89	0.64
	SD	3.74	5.94	2.93	1.59	1.29	1.89	1.77	0.62
SacchCONF (56)	MAE	4.13	4.71	3.08	1.52	0.76	1.17	1.68	0.54
	MSE	1.67	2.74	0.81	-0.63	-0.13	0.94	-1.00	0.18
	MAXE	18.84	21.05	12.77	4.97	2.07	4.15	8.07	2.11
	RMSE	6.59	7.49	4.62	1.88	0.94	1.47	2.63	0.69
	SD	6.43	7.03	4.59	1.78	0.94	1.15	2.45	0.68
CCONF (426)	MAE	4.59	5.28	3.79	1.46	1.26	1.20	2.02	0.87

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
	MSE	-1.23	-1.34	-1.51	-0.75	0.82	0.71	0.54	-0.57
	MAXE	15.42	17.62	10.90	6.64	7.48	4.56	8.07	3.77
	RMSE	5.91	6.84	4.62	1.94	1.63	1.53	2.62	1.07
	SD	5.79	6.71	4.37	1.79	1.40	1.36	2.56	0.91
ACONF (15)	MAE	0.19	1.44	0.89	0.37	1.33	0.18	0.37	0.40
	MSE	0.06	-1.44	-0.89	-0.37	1.33	-0.18	0.37	-0.40
	MAXE	0.70	2.63	1.76	0.74	2.67	0.60	1.11	0.75
	RMSE	0.26	1.55	0.96	0.40	1.49	0.24	0.47	0.43
	SD	0.26	0.58	0.36	0.17	0.70	0.17	0.30	0.15
BCONF (64)	MAE	1.21	2.40	0.58	0.28	0.26	1.25	0.64	0.22
	MSE	1.21	2.34	0.53	0.09	0.05	1.17	-0.64	-0.12
	MAXE	2.22	3.69	1.40	1.14	1.08	1.81	0.99	0.62
	RMSE	1.30	2.53	0.69	0.36	0.32	1.30	0.68	0.28
	SD	0.48	0.98	0.44	0.35	0.32	0.57	0.25	0.25
PentCONF (342)	MAE	0.39	0.96	0.55	0.25	1.05	0.47	0.88	0.16
	MSE	-0.37	-0.96	-0.54	-0.11	1.05	0.46	0.88	-0.10
	MAXE	1.23	1.94	1.27	0.86	2.25	1.77	2.55	0.47
	RMSE	0.50	1.07	0.65	0.30	1.15	0.58	1.00	0.19
	SD	0.34	0.49	0.36	0.28	0.48	0.36	0.47	0.16
Undecamer125 (124)	MAE	1.67	1.57	1.27	0.45	0.56	2.02	0.84	0.34
	MSE	0.14	1.57	0.32	-0.15	-0.50	0.93	-0.32	-0.17
	MAXE	4.85	3.43	3.87	1.72	2.25	4.89	3.39	1.47
	RMSE	1.95	1.69	1.54	0.57	0.74	2.31	1.03	0.46
	SD	1.95	0.63	1.51	0.56	0.54	2.12	0.98	0.43
ICONF (17)	MAE	1.76	1.74	2.29	1.00	1.00	1.00	1.22	0.39
	MSE	-1.29	-1.12	-1.39	-0.38	0.13	0.30	0.02	0.10
	MAXE	7.39	5.65	8.28	3.28	2.35	3.11	3.08	0.89
	RMSE	2.91	2.55	3.43	1.34	1.20	1.37	1.63	0.47
	SD	2.69	2.36	3.22	1.33	1.23	1.38	1.68	0.47
MCONF (51)	MAE	2.66	0.88	0.89	1.04	2.28	1.05	0.30	0.48
	MSE	-2.66	0.44	-0.34	-0.97	-2.05	1.04	0.26	-0.18
	MAXE	4.95	2.97	2.15	2.01	4.49	1.94	0.90	1.12
	RMSE	3.05	1.19	1.10	1.20	2.61	1.17	0.36	0.58
	SD	1.51	1.12	1.06	0.71	1.64	0.52	0.25	0.56
Torsion21 (189)	MAE	1.07	1.30	1.17	0.53	0.62	0.80	0.89	0.22
	MSE	0.26	0.42	0.53	-0.38	0.57	0.72	0.83	0.00
	MAXE	3.13	4.39	3.53	2.00	2.15	2.47	3.36	0.98
	RMSE	1.33	1.65	1.50	0.71	0.80	0.97	1.19	0.29
	SD	1.31	1.59	1.40	0.60	0.56	0.64	0.85	0.29
37Conf8 (258)	MAE	2.12	2.48	1.90	1.55	2.12	1.13	1.56	0.90
	MSE	-0.98	-0.77	-1.29	-0.90	0.28	0.49	-0.04	-0.34
	MAXE	11.66	16.36	10.70	9.43	9.28	7.77	8.38	5.48
	RMSE	2.88	3.49	2.67	2.24	2.70	1.61	2.14	1.33
	SD	2.71	3.42	2.34	2.05	2.69	1.53	2.15	1.29

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
DCONF (2142)	MAE	0.77	1.05	0.88	0.58	0.48	0.59	0.50	0.30
	MSE	0.09	0.34	0.22	-0.08	0.27	0.52	0.40	0.13
	MAXE	3.40	3.53	4.17	3.59	2.28	2.46	2.63	1.39
	RMSE	1.01	1.34	1.16	0.82	0.65	0.81	0.68	0.42
	SD	1.01	1.30	1.13	0.82	0.59	0.62	0.55	0.40
MolCONF (5623)	MAE	1.01	0.76	0.56	0.49	0.96	0.42	0.45	0.35
	MSE	-0.42	-0.03	-0.21	-0.27	-0.31	0.08	-0.10	-0.10
	MAXE	18.17	9.87	10.13	16.43	20.87	5.79	6.17	6.28
	RMSE	2.02	1.28	0.98	0.93	2.05	0.70	0.78	0.58
	SD	1.98	1.28	0.96	0.88	2.02	0.70	0.77	0.57
MOLdef (9298)	MAE	3.90	3.82	3.32	2.36	1.91	2.05	3.41	1.49
	MSE	2.77	2.83	2.89	0.78	1.69	1.75	1.81	1.12
	MAXE	79.81	79.65	73.23	54.38	29.89	29.73	69.13	31.70
	RMSE	6.08	5.94	4.79	4.05	2.77	2.95	6.20	2.33
	SD	5.42	5.22	3.82	3.97	2.19	2.38	5.93	2.04
MOLdef-H2O (990)	MAE	1.77	1.78	1.02	0.37	0.60	0.62	1.37	0.49
	MSE	0.03	0.04	0.24	-0.05	0.38	0.39	0.59	-0.26
	MAXE	10.07	10.06	5.82	4.19	2.90	2.95	6.73	3.97
	RMSE	2.57	2.58	1.40	0.61	0.88	0.92	2.10	0.70
	SD	2.57	2.58	1.38	0.61	0.80	0.83	2.02	0.64
ANI1ccxCONF (32944)	MAE	5.93	5.86	4.44	2.82	3.76	3.96	7.08	1.97
	MSE	2.57	2.70	3.74	0.96	3.38	3.51	4.55	1.01
	MAXE	71.00	69.05	38.86	39.10	36.17	38.12	99.51	19.74
	RMSE	8.47	8.35	6.34	4.13	5.23	5.47	10.33	2.79
	SD	8.08	7.90	5.12	4.02	3.99	4.20	9.28	2.60
PEPCONF-Dipeptide (anionic) (175)	MAE	2.31	2.37	1.98	1.11	1.10	1.26	1.03	0.85
	MSE	-1.05	-0.34	-0.49	-0.38	-0.21	0.50	0.36	-0.08
	MAXE	7.15	7.97	5.69	3.55	5.56	4.08	3.42	2.66
	RMSE	2.82	2.89	2.37	1.38	1.43	1.60	1.28	1.03
	SD	2.63	2.88	2.32	1.34	1.42	1.52	1.23	1.03
MolCONF (anionic) (79)	MAE	2.89	2.76	2.12	2.11	0.64	0.68	2.57	0.48
	MSE	2.59	2.49	0.61	0.19	-0.48	-0.58	-2.46	-0.04
	MAXE	17.70	16.43	13.57	10.61	3.23	2.46	13.25	2.38
	RMSE	5.84	5.49	4.41	4.02	1.09	1.11	5.00	0.82
	SD	5.27	4.93	4.40	4.05	0.99	0.95	4.38	0.82

Table S2. The error analysis with respect to reference data of the various datasets present in the validation set. The numbers in bracket in the first column indicates the number of datapoints. MAE = mean absolute error in kcal/mol, MSE = mean signed error in kcal/mol, MAXE = maximum absolute error in kcal/mol, RMSE = root-mean- square error in kcal/mol, SD = standard deviation in kcal/mol.

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
BlindNCI (80)	MAE	1.29	1.00	0.38	0.34	1.29	1.10	0.50	0.26
	MSE	1.21	-0.97	0.09	0.03	1.09	-1.10	-0.04	0.09
	MAXE	15.36	7.82	3.44	4.02	19.56	6.98	4.28	2.87
	RMSE	2.91	2.03	0.83	0.64	3.20	1.89	0.90	0.51
	SD	2.66	1.80	0.83	0.64	3.03	1.55	0.91	0.51
DES15K (11474)	MAE	4.83	2.66	2.14	1.36	5.13	1.80	2.34	0.60
	MSE	4.48	-2.24	1.37	0.26	5.07	-1.65	1.96	0.05
	MAXE	28.49	21.86	26.11	16.78	27.69	8.57	24.34	7.56
	RMSE	6.72	3.91	4.02	2.36	7.33	2.17	3.93	0.94
	SD	5.01	3.20	3.78	2.34	5.30	1.41	3.41	0.94
NENCI-2021 (5859)	MAE	3.86	2.26	1.61	1.09	3.93	1.86	1.80	0.50
	MSE	3.47	-1.84	0.93	0.56	3.80	-1.51	1.25	0.12
	MAXE	50.14	21.45	39.84	37.25	41.85	12.55	26.86	10.65
	RMSE	6.56	3.26	3.71	2.45	7.15	2.29	3.84	0.95
	SD	5.57	2.70	3.59	2.39	6.06	1.72	3.63	0.94
R160x6 (960)	MAE	1.55	0.81	0.67	0.84	2.02	0.70	1.03	0.83
	MSE	1.42	-0.57	0.30	0.32	1.98	-0.01	0.86	0.47
	MAXE	15.58	6.20	9.57	7.54	18.27	6.26	8.90	7.68
	RMSE	2.40	1.12	1.10	1.23	2.97	1.00	1.50	1.22
	SD	1.93	0.96	1.06	1.19	2.22	1.00	1.24	1.13
R739x5 (4330)	MAE	1.73	0.72	0.90	0.64	1.86	0.56	0.89	0.50
	MSE	1.71	-0.27	0.72	0.14	1.85	-0.13	0.86	0.14
	MAXE	19.06	13.65	21.27	8.24	9.99	4.18	11.13	3.91
	RMSE	2.34	1.20	1.82	0.90	2.20	0.77	1.26	0.67
	SD	1.60	1.17	1.68	0.89	1.19	0.76	0.93	0.65
CE20 (20)	MAE	7.90	16.64	3.32	2.37	2.15	11.10	5.08	1.95
	MSE	7.13	16.64	2.33	0.39	1.59	11.10	-3.21	1.61
	MAXE	31.78	43.29	9.63	8.22	5.84	22.33	26.68	5.54
	RMSE	11.70	20.39	4.16	3.16	2.87	12.51	8.74	2.66
	SD	9.51	12.09	3.53	3.22	2.45	5.93	8.34	2.17
CHAL336 (48)	MAE	4.22	2.17	1.30	2.20	3.88	2.23	0.91	2.16
	MSE	4.22	-1.78	0.46	-2.19	3.78	-2.23	0.01	-2.16
	MAXE	16.58	11.47	11.65	19.62	11.24	14.07	6.31	11.82
	RMSE	5.44	2.80	2.67	3.60	4.59	2.91	1.58	2.85
	SD	3.47	2.19	2.66	2.89	2.63	1.89	1.59	1.88
XB45 (33)	MAE	8.43	6.66	7.33	2.79	3.18	2.01	1.98	2.78
	MSE	-7.70	-3.19	-6.04	0.52	-2.98	1.53	-1.32	2.61
	MAXE	42.26	32.32	38.25	10.12	12.58	6.39	7.90	6.78
	RMSE	15.59	11.51	13.62	3.74	4.31	2.67	2.65	3.27

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
	SD	13.77	11.23	12.40	3.76	3.16	2.22	2.34	1.99
WaterOrg (2376)	MAE	1.83	1.81	0.71	0.79	1.18	2.44	0.54	0.40
	MSE	1.82	-1.80	0.69	0.77	1.18	-2.44	0.06	0.25
	MAXE	4.86	4.57	2.86	2.65	3.03	5.97	2.41	1.26
	RMSE	2.00	2.03	0.93	0.93	1.29	2.63	0.63	0.47
	SD	0.84	0.94	0.62	0.52	0.52	0.97	0.62	0.40
Water27 (27)	MAE	26.90	49.77	21.88	14.38	7.13	29.93	7.42	7.00
	MSE	26.90	49.77	21.88	14.14	7.06	29.93	2.04	5.47
	MAXE	51.43	119.82	46.79	41.27	23.45	92.86	28.14	24.41
	RMSE	31.45	60.29	26.34	20.52	10.47	39.00	10.55	10.75
	SD	16.61	34.68	14.95	15.16	7.89	25.47	10.55	9.43
HW6Cl (6)	MAE	2.32	12.20	4.03	2.90	0.72	14.57	1.66	0.59
	MSE	2.32	-12.20	4.03	-2.90	-0.06	-14.57	1.66	-0.05
	MAXE	3.30	22.83	5.32	4.04	1.03	26.31	2.72	0.84
	RMSE	2.60	14.09	4.27	3.10	0.75	16.94	1.79	0.61
	SD	1.29	7.74	1.53	1.19	0.82	9.47	0.73	0.67
HW6F (6)	MAE	49.64	61.61	44.13	30.29	26.11	38.08	20.60	24.52
	MSE	-49.64	-61.61	-44.13	-30.29	-26.11	-38.08	-20.60	-24.52
	MAXE	57.78	81.32	50.52	35.83	33.91	57.45	26.65	30.70
	RMSE	50.13	63.23	44.46	30.71	27.21	40.94	21.47	25.29
	SD	7.68	15.62	5.90	5.50	8.39	16.47	6.62	6.80
FmH2O10 (10)	MAE	59.30	107.13	50.73	35.86	37.74	85.57	29.17	33.32
	MSE	-59.30	-107.13	-50.73	-35.86	-37.74	-85.57	-29.17	-33.32
	MAXE	62.39	109.19	53.26	37.22	38.66	86.77	30.53	34.55
	RMSE	59.33	107.14	50.75	35.88	37.74	85.58	29.17	33.34
	SD	1.81	0.93	1.65	1.29	0.47	1.30	0.66	0.91
SW49Bind345 (30)	MAE	4.58	11.36	4.66	3.90	0.77	7.57	0.86	1.72
	MSE	-4.48	-11.30	-4.59	-3.39	-0.67	-7.48	-0.78	-1.72
	MAXE	8.53	20.56	8.79	10.08	1.54	13.48	2.42	3.11
	RMSE	5.50	13.69	5.61	4.72	0.90	9.08	1.05	1.95
	SD	3.24	7.86	3.28	3.34	0.60	5.22	0.71	0.94
SW49Bind6 (18)	MAE	11.13	27.96	11.20	8.93	2.02	18.85	2.09	3.66
	MSE	-11.13	-27.96	-11.20	-8.93	-2.02	-18.85	-2.09	-3.66
	MAXE	11.63	29.23	12.62	14.68	2.28	19.79	3.79	4.27
	RMSE	11.14	27.99	11.22	9.15	2.05	18.88	2.25	3.68
	SD	0.23	1.15	0.75	2.02	0.31	0.96	0.86	0.42
H2O20Bind10 (10)	MAE	18.91	110.97	16.90	5.11	11.48	103.54	9.47	2.45
	MSE	-18.91	-110.97	-16.90	-5.11	-11.48	-103.54	-9.47	-2.45
	MAXE	20.10	112.79	17.73	5.60	13.13	110.41	11.89	3.98
	RMSE	18.96	110.98	16.93	5.12	11.49	103.57	9.51	2.54
	SD	1.40	1.12	1.06	0.26	0.68	2.72	0.92	0.70
L7 (7)	MAE	20.60	3.64	1.37	2.97	21.22	3.61	2.26	2.20
	MSE	20.60	-3.64	0.50	2.44	21.22	-3.02	1.12	1.46
	MAXE	39.89	6.82	2.69	6.94	40.72	10.39	5.17	4.41

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
	RMSE	23.60	4.38	1.54	3.63	24.68	4.73	2.83	2.60
	SD	12.43	2.63	1.57	2.91	13.61	3.93	2.81	2.32
S12L (10)	MAE	25.23	14.65	6.05	3.76	26.37	13.51	4.86	2.91
	MSE	25.23	-14.65	-4.68	-2.15	26.37	-13.51	-3.54	-2.39
	MAXE	49.53	25.18	11.75	13.01	56.26	21.01	8.09	8.65
	RMSE	29.16	15.84	7.55	5.95	30.53	14.58	5.42	3.79
	SD	15.40	6.36	6.24	5.85	16.22	5.77	4.33	3.10
S30L (26)	MAE	33.41	13.25	4.80	4.20	34.95	11.71	3.75	3.24
	MSE	33.41	-13.25	-3.26	0.65	34.95	-11.71	-1.72	0.65
	MAXE	68.05	41.59	10.23	12.17	73.13	38.15	13.05	9.11
	RMSE	37.99	15.05	5.75	5.35	39.73	13.52	4.96	4.11
	SD	18.44	7.29	4.84	5.42	19.27	6.90	4.74	4.14
C60dimer (14)	MAE	8.83	1.91	0.90	2.19	10.62	1.13	1.78	1.62
	MSE	8.82	-1.91	-0.90	2.18	10.61	-0.12	0.89	1.61
	MAXE	20.23	7.61	3.04	4.10	30.53	2.86	7.26	3.93
	RMSE	10.59	2.64	1.10	2.38	13.71	1.36	2.71	1.86
	SD	6.09	1.90	0.67	1.00	9.00	1.41	2.65	0.96
Ni2021 (11)	MAE	41.64	25.53	6.74	14.28	44.96	22.01	9.71	19.05
	MSE	41.64	-23.69	5.22	14.28	44.96	-20.37	8.54	19.05
	MAXE	122.94	99.45	21.05	50.28	165.06	57.33	46.90	77.59
	RMSE	51.46	35.81	8.93	19.07	61.17	26.05	16.31	27.82
	SD	31.71	28.16	7.59	13.25	43.50	17.04	14.57	21.27
Anionpi (anionic) (16)	MAE	7.25	8.17	6.96	4.49	4.91	5.56	3.55	3.79
	MSE	-1.12	-7.30	-3.59	-2.10	1.08	-5.09	-1.39	-2.78
	MAXE	34.31	48.58	40.33	28.35	13.46	19.08	12.17	12.98
	RMSE	10.58	14.67	11.81	8.38	5.59	7.91	4.87	5.37
	SD	10.86	13.15	11.62	8.38	5.67	6.25	4.82	4.75
IL236 (anionic) (236)	MAE	4.74	6.78	3.13	2.11	4.59	4.71	1.97	1.50
	MSE	2.34	-6.76	-0.60	-0.30	4.47	-4.63	1.53	-1.21
	MAXE	10.56	17.24	11.55	7.09	10.62	8.79	8.60	5.85
	RMSE	5.42	7.90	4.01	2.60	5.24	5.36	2.87	1.90
	SD	4.90	4.09	3.97	2.59	2.75	2.70	2.44	1.46
DES15K (anionic) (1281)	MAE	6.53	7.40	6.89	4.30	5.05	4.07	4.92	2.63
	MSE	0.06	-5.70	-0.49	-1.20	2.74	-3.03	2.18	-2.04
	MAXE	65.05	68.00	67.57	57.81	20.76	24.52	27.22	28.97
	RMSE	10.74	12.15	11.23	7.79	7.06	5.66	7.37	4.84
	SD	10.74	10.74	11.23	7.70	6.51	4.79	7.04	4.39
NENCI-2021 (anionic) (889)	MAE	8.01	8.25	7.20	6.18	5.00	5.17	4.19	3.38
	MSE	-0.21	-4.36	-1.37	-0.41	1.79	-2.35	0.63	-1.62
	MAXE	57.60	62.79	64.39	63.40	33.50	22.42	28.76	19.97
	RMSE	11.63	12.15	10.97	9.58	6.96	6.40	5.72	4.72
	SD	11.64	11.35	10.89	9.58	6.73	5.96	5.69	4.43
CHAL336 (anionic) (19)	MAE	15.48	15.66	14.10	16.46	6.89	5.86	5.44	6.82
	MSE	-11.05	-15.23	-12.57	-15.96	-0.30	-4.48	-1.82	-6.47

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
	MAXE	57.03	61.20	57.85	58.91	14.69	18.86	15.51	24.66
	RMSE	21.97	23.90	21.90	24.61	7.81	8.24	6.79	10.25
	SD	19.51	18.93	18.42	19.25	8.02	7.10	6.73	8.17
XB45 (anionic) (12)	MAE	30.25	33.13	31.40	30.22	13.13	15.64	14.25	14.74
	MSE	29.44	33.13	30.98	30.22	11.87	15.57	13.42	14.45
	MAXE	84.76	88.80	86.47	76.31	39.19	40.73	37.67	35.00
	RMSE	43.82	46.02	45.07	40.31	20.16	22.09	21.14	19.47
	SD	33.89	33.36	34.19	27.86	17.02	16.37	17.06	13.63
S30L (anionic) (2)	MAE	4.67	31.48	15.50	6.09	12.21	14.60	1.38	3.97
	MSE	-4.67	-31.48	-15.50	-6.09	12.21	-14.60	1.38	-3.97
	MAXE	6.96	31.57	15.83	6.47	14.65	14.67	1.56	4.83
	RMSE	5.20	31.48	15.51	6.10	12.45	14.60	1.39	4.06
	SD	3.24	0.12	0.46	0.55	3.45	0.09	0.25	1.22
SafroleCONF (5)	MAE	0.59	0.92	0.82	0.70	0.69	0.44	0.39	0.45
	MSE	-0.59	-0.92	-0.82	-0.70	-0.11	-0.44	-0.35	-0.45
	MAXE	1.32	1.28	1.19	1.19	1.01	0.97	0.89	1.05
	RMSE	0.82	1.03	0.92	0.82	0.77	0.62	0.55	0.65
	SD	0.65	0.52	0.47	0.48	0.85	0.48	0.48	0.53
AlcoholCONF (31)	MAE	0.66	0.71	0.61	0.40	0.33	0.35	0.47	0.27
	MSE	-0.48	-0.48	-0.37	-0.06	0.31	0.31	0.42	0.11
	MAXE	2.20	2.30	2.05	0.94	0.69	0.90	1.08	0.66
	RMSE	0.90	0.89	0.77	0.48	0.37	0.41	0.55	0.33
	SD	0.77	0.76	0.68	0.48	0.22	0.28	0.35	0.31
BeranCONF (50)	MAE	1.47	1.54	1.03	0.80	0.81	0.65	1.27	0.33
	MSE	0.21	0.33	0.47	-0.06	0.35	0.47	0.61	-0.01
	MAXE	6.03	5.33	3.88	3.77	2.71	2.64	7.08	1.38
	RMSE	1.91	2.09	1.35	1.12	1.09	0.85	1.81	0.45
	SD	1.91	2.09	1.28	1.13	1.04	0.72	1.72	0.45
Torsion30 (2107)	MAE	0.92	1.62	1.18	0.54	0.59	0.59	0.63	0.43
	MSE	0.38	0.65	0.66	0.30	0.17	0.44	0.44	0.24
	MAXE	8.25	7.98	8.58	11.56	13.81	11.50	14.14	11.11
	RMSE	1.24	2.03	1.54	0.94	1.07	1.03	1.19	0.81
	SD	1.18	1.92	1.40	0.89	1.06	0.93	1.11	0.78
MPCONF196 (112)	MAE	3.61	3.90	2.86	2.01	3.05	3.51	2.76	1.08
	MSE	-2.77	-0.23	-1.82	-0.62	0.57	3.11	1.52	0.37
	MAXE	12.57	16.14	11.19	5.68	11.10	12.23	10.97	4.56
	RMSE	4.52	5.01	3.63	2.40	3.87	4.34	3.48	1.49
	SD	3.59	5.03	3.15	2.32	3.85	3.05	3.14	1.45
PEPCONF-Tripeptide (647)	MAE	3.05	2.26	1.33	1.18	2.51	1.94	1.40	0.80
	MSE	-1.97	1.19	-0.54	-0.44	-1.58	1.58	-0.15	0.05
	MAXE	11.27	10.01	6.25	4.76	11.02	7.18	6.46	3.42
	RMSE	3.84	2.83	1.70	1.50	3.17	2.42	1.82	1.00
	SD	3.30	2.58	1.62	1.43	2.75	1.84	1.81	1.00
	MAE	5.23	3.71	2.64	2.75	5.61	2.53	2.76	2.67

Dataset (#)	Error measure	HF/MINIX	HF-D3/MINIX	HF-3c	HF/MINIX-ACP	HF/6-31G*	HF-D3/6-31G*	HF/6-31G* with 3c	HF/6-31G*-ACP
PEPCONF-Disulfide (620)	MSE	-2.99	1.84	-0.83	-1.68	-3.63	1.20	-1.47	-1.69
	MAXE	30.74	16.00	16.40	17.01	33.17	9.56	17.25	16.92
	RMSE	6.97	4.71	3.56	3.74	7.39	3.15	3.83	3.80
	SD	6.30	4.34	3.46	3.34	6.45	2.92	3.54	3.40
PEPCONF-Cyclic (320)	MAE	3.87	3.18	3.23	1.87	4.92	4.62	4.65	1.31
	MSE	-2.11	-1.14	-1.65	0.08	3.58	4.54	4.03	0.16
	MAXE	16.63	15.38	12.49	6.98	18.52	14.47	20.68	6.17
	RMSE	5.02	4.18	4.07	2.39	6.18	5.35	5.78	1.67
	SD	4.57	4.02	3.73	2.40	5.05	2.83	4.15	1.67
PEPCONF-Bioactive (175)	MAE	4.21	3.58	2.46	1.91	3.57	1.82	1.49	0.99
	MSE	-1.31	2.02	0.30	0.03	-1.93	1.41	-0.32	0.01
	MAXE	20.34	14.50	10.22	7.12	16.91	6.99	5.77	3.84
	RMSE	5.58	4.58	3.19	2.44	4.73	2.40	1.90	1.25
	SD	5.44	4.12	3.19	2.45	4.33	1.95	1.88	1.25
PEPCONF-Disulfide (anionic) (150)	MAE	8.44	3.96	5.02	4.78	8.00	2.16	4.35	4.44
	MSE	-7.53	-1.91	-4.33	-4.23	-6.81	-1.20	-3.62	-3.96
	MAXE	31.85	18.97	21.68	17.06	35.48	7.33	18.50	18.49
	RMSE	11.25	5.30	6.80	6.39	10.52	2.73	5.87	5.86
	SD	8.38	4.96	5.25	4.81	8.05	2.47	4.63	4.33
PEPCONF-Bioactive (anionic) (20)	MAE	4.98	2.99	2.70	2.37	2.78	1.40	1.04	0.79
	MSE	-0.77	0.56	-0.59	-1.07	-0.39	0.93	-0.22	-0.41
	MAXE	10.85	11.00	9.60	5.86	7.57	4.14	2.59	2.51
	RMSE	5.86	4.16	3.51	2.94	3.56	1.76	1.34	1.01
	SD	5.96	4.23	3.55	2.81	3.63	1.53	1.35	0.95

Table S3. Comparison of percentage change in the single-point (SP) calculation time between that of uncorrected and ACP corrected approaches for the S30L data set. The shorthand notations used are as follows: MINIX = [((SP time of HF/MINIX-ACP) - (SP time of HF/MINIX)) / (SP time of HF/MINIX) X 100%] and 6-31G* = [((SP time of HF/6-31G*-ACP) - (SP time of HF/6-31G*)) / (SP time of HF/6-31G*) X 100%]. Single-point calculations were performed using Gaussian16 package and 32 cores of Dell EMC R440 CPU compute nodes on *Sockeye* cluster (University of British Columbia's Advanced Research Computing facility).

Molecule	Atoms	MINIX (in %)	6-31G* (in %)
S30L_001_dimer	92	28.6	-2.9
S30L_002_dimer	86	25.5	-2.7
S30L_003_dimer	126	15.3	-2.5
S30L_004_dimer	113	19.5	0.9
S30L_005_dimer	100	2.9	-3.7
S30L_006_dimer	92	30.1	-3.3
S30L_007_dimer	156	20.3	-2.9
S30L_008_dimer	180	19.9	4.9
S30L_009_dimer	148	15.3	-7.1
S30L_010_dimer	158	25.4	3.9

Molecule	Atoms	MINIX (in %)	6-31G* (in %)
S30L_011_dimer	140	19.6	-4.7
S30L_012_dimer	150	20.1	0.0
S30L_013_dimer	205	29.0	3.5
S30L_014_dimer	204	16.4	4.0
S30L_017_dimer	144	13.7	-58.0
S30L_018_dimer	142	13.2	5.5
S30L_019_dimer	163	18.4	6.2
S30L_020_dimer	172	27.0	6.9
S30L_021_dimer	153	12.7	6.0
S30L_022_dimer	133	16.0	9.2
S30L_023_dimer	98	14.9	5.6
S30L_024_dimer	184	31.8	11.9
S30L_025_dimer	142	26.8	12.5
S30L_026_dimer	142	23.4	15.3
S30L_027_dimer	125	15.4	5.4
S30L_028_dimer	122	14.7	12.8
S30L_029_dimer	121	9.8	11.9
S30L_030_dimer	128	10.5	11.8

Appendix 5

Supporting Information for Chapter 8

Section S1. Sample input file demonstrating the use of atom-centered potentials in Gaussian16 software

The MINIS, MINIX, and 6-31G* basis-set files (in .gbs extension) and the corresponding ACP files (in .acp extension) are provided separately in the supporting information ZIP file accompanying this document. An externally specified basis set file named **“minis.gbs”** and the additional ACP file **“minis.acp”** is defined and invoked by adding the keyword **“genECP”** to the route section of the Gaussian input file. Note that the ACP are not transferable and are proposed to be used with their underlying methods only. The D3 dispersion correction parameters for HF method should also be defined before performing calculations. This is done using the “IOp” option of Gaussian as shown below. The *IOp* option for D3 correction to be defined are 3/174, 3/175, 3/177, and 3/178. ACP were developed for use with D3 dispersion correction parameters that correspond to those for the HF/aug-cc-pVTZ method and with Becke-Johnson damping i.e. $s_6 = 1.0$, $s_8 = 0.9171$, $a_1(BJ) = 0.3385$, and $a_2(BJ) = 2.8830$ Å. Therefore, the values of *IOp* to be defined in the Gaussian route section are 3/174=1000000, 3/175=917100, 3/177=338500, and 3/178=2883000.

```
%mem=4GB
%nprocs=8
# HF empiricaldispersion=gd3bj genECP IOp(3/174=1000000,3/175=917100,3/177=338500,3/178=2883000)

Title: Sample water dimer input using HF-D3/MINIS-ACP method

0 1
O   -0.702196054   -0.056060256    0.009942262
H   -1.022193224    0.846775782   -0.011488714
H    0.257521062    0.042121496    0.005218999
O    2.220871067    0.026716792    0.000620476
H    2.597492682   -0.411663274    0.766744858
H    2.593135384   -0.449496183   -0.744782026

@minis.gbs/N

@minis.acp/N
```

Section S2. MINIs basis set file

```
-H     0
S   3   1.00
      7.034063              0.070452
      1.064756              0.407826
      0.236559              0.647752
****
-B     0
S   3   1.00
      4.457854             -0.082419
      0.369315              0.559064
      0.122555              0.516795
S   3   1.00
    108.43704               0.068651
     16.120560              0.389933
      3.3734300             0.671395
P   3   1.00
      3.214892              0.105900
      0.646136              0.457180
      0.153916              0.631861
****
-C     0
S   3   1.00
      6.616612             -0.081380
      0.525856              0.574853
      0.169958              0.502413
S   3   1.00
    153.17226               0.070740
     23.073030              0.395380
      4.9232900             0.663311
P   3   1.00
      4.912920              0.109931
      0.997616              0.462713
      0.232685              0.627514
****
-N     0
S   3   1.00
      8.919426             -0.080890
      0.706141              0.567202
      0.225054              0.511092
S   3   1.00
    218.36449               0.067870
     32.598890              0.390202
      6.9173900             0.670083
P   3   1.00
      6.556272              0.115919
      1.349079              0.469958
      0.312209              0.618448
****
-O     0
S   3   1.00
     11.789326             -0.080820
      0.9128940             0.582090
      0.2866610             0.497160
S   3   1.00
    281.86658               0.069060
     42.416000              0.393159
      9.0956200             0.665669
P   3   1.00
      8.274140              0.124271
      1.715463              0.476594
      0.383013              0.613044
****
```

```
-F     0
S   3   1.00
    368.37112                0.067040
     55.061060               0.389249
     11.747670               0.670788
S   3   1.00
     15.364708              -0.080550
      1.1675460              0.587729
      0.3631410              0.491979
P   3   1.00
     10.725667               0.126270
      2.2258170              0.477948
      0.4861050              0.614008
****
-Si     0
S   3   1.00
    909.23487                0.066405
    137.12456                0.386222
     29.714810               0.672240
S   3   1.00
     39.129423              -0.090999
      3.335981              0.611615
      1.251259              0.456860
S   3   1.00
      2.197649             -0.168733
      0.275927              0.675453
      0.100425              0.429419
P   3   1.00
     37.881761               0.108753
      8.304598              0.463515
      2.120792              0.611334
P   3   1.00
      0.545789              0.238913
      0.208220              0.542295
      0.076007              0.345453
****
-P     0
S   3   1.00
     45.450377              -0.092655
      3.899926              0.626513
      1.488507              0.441039
S   3   1.00
   1053.2658                 0.065865
    158.79044                0.384578
     34.424407               0.673963
S   3   1.00
      2.469483             -0.180549
      0.320872              0.680952
      0.116832              0.429142
P   3   1.00
     46.100019               0.105388
     10.165057               0.459712
      2.644794              0.613714
P   3   1.00
      0.679059              0.235885
      0.257826              0.554160
      0.092783              0.336530
****
-S     0
S   3   1.00
     52.13903              -0.094232
      4.528799              0.635468
      1.754938              0.431506
S   3   1.00
```

```
  1201.4584           0.065765
   181.39212          0.383948
    39.404795         0.674372
S   3   1.00
     2.920526        0.190042
     0.392187       -0.685527
     0.142699       -0.429272
P   3   1.00
    54.644071         0.103673
    12.122902         0.458190
     3.206504        0.613400
P   3   1.00
     0.887615        0.229436
     0.327100        0.552960
     0.111743        0.353700
****
-Cl     0
S   3   1.00
    59.225732        -0.095620
     5.213902        0.641426
     2.047346        0.425153
S   3   1.00
  1362.0220           0.065544
   205.81110          0.382987
    44.772167         0.675210
S   3   1.00
     3.447124        0.196401
     0.473785       -0.692360
     0.171321       -0.426193
P   3   1.00
    64.099958         0.101789
    14.287139         0.456107
     3.828135        0.614282
P   3   1.00
     1.103904        0.235903
     0.399178        0.558066
     0.133236        0.346600
****
```

Section S3. MINIX basis set file

```
-H     0
S   3   1.00
     7.034063        0.070452
     1.064756        0.407826
     0.236559        0.647752
****
-B     0
S   3   1.00
     4.457854       -0.082419
     0.369315        0.559064
     0.122555        0.516795
S   3   1.00
   108.43704          0.068651
    16.120560         0.389933
     3.3734300        0.671395
P   3   1.00
     3.214892        0.105900
     0.646136        0.457180
     0.153916        0.631861
****
-C     0
S   3   1.00
```

```
      6.616612              -0.081380
      0.525856               0.574853
      0.169958               0.502413
S   3   1.00
    153.17226                0.070740
     23.073030               0.395380
      4.9232900              0.663311
P   3   1.00
      4.912920               0.109931
      0.997616               0.462713
      0.232685               0.627514
****
-N     0
S   3   1.00
      8.919426              -0.080890
      0.706141               0.567202
      0.225054               0.511092
S   3   1.00
    218.36449                0.067870
     32.598890               0.390202
      6.9173900              0.670083
P   3   1.00
      6.556272               0.115919
      1.349079               0.469958
      0.312209               0.618448
****
-O     0
S   3   1.00
     11.789326              -0.080820
      0.9128940              0.582090
      0.2866610              0.497160
S   3   1.00
    281.86658                0.069060
     42.416000               0.393159
      9.0956200              0.665669
P   3   1.00
      8.274140               0.124271
      1.715463               0.476594
      0.383013               0.613044
****
-F     0
S   3   1.00
    368.37112                0.067040
     55.061060               0.389249
     11.747670               0.670788
S   3   1.00
     15.364708              -0.080550
      1.1675460              0.587729
      0.3631410              0.491979
P   3   1.00
     10.725667               0.126270
      2.2258170              0.477948
      0.4861050              0.614008
****
-Si     0
S   3   1.00
    909.2348700              0.0664050
    137.1245600              0.3862220
     29.7148100              0.6722400
S   3   1.00
     39.1294230             -0.0909990
      3.3359810              0.6116150
      1.2512590              0.4568600
S   3   1.00
```

```
      2.1976490             -0.1687330
      0.2759270              0.6754530
      0.1004250              0.4294190
P   3   1.00
     37.8817610              0.1087530
      8.3045980              0.4635150
      2.1207920              0.6113340
P   3   1.00
      0.5457890              0.2389130
      0.2082200              0.5422950
      0.0760070              0.3454530
D   1   1.00
      0.3500000              1.0000000
****
-P     0
S   3   1.00
   1053.2658000              0.0658650
    158.7904400              0.3845780
     34.4244070              0.6739630
S   3   1.00
     45.4503770             -0.0926550
      3.8999260              0.6265130
      1.4885070              0.4410390
S   3   1.00
      2.4694830             -0.1805490
      0.3208720              0.6809520
      0.1168320              0.4291420
P   3   1.00
     46.1000190              0.1053880
     10.1650570              0.4597120
      2.6447940              0.6137140
P   3   1.00
      0.6790590              0.2358850
      0.2578260              0.5541600
      0.0927830              0.3365300
D   1   1.00
      0.4500000              1.0000000
****
-S     0
S   3   1.00
   1201.4584000              0.0657650
    181.3921200              0.3839480
     39.4047950              0.6743720
S   3   1.00
     52.1390300             -0.0942320
      4.5287990              0.6354680
      1.7549380              0.4315060
S   3   1.00
      2.9205260              0.1900420
      0.3921870             -0.6855270
      0.1426990             -0.4292720
P   3   1.00
     54.6440710              0.1036730
     12.1229020              0.4581900
      3.2065040              0.6134000
P   3   1.00
      0.8876150              0.2294360
      0.1117430              0.3537000
      0.3271000              0.5529600
D   1   1.00
      0.5500000              1.0000000
****
-Cl     0
S   3   1.00
```

```
  1362.0220000          0.0655440
   205.8111000          0.3829870
    44.7721670          0.6752100
S   3   1.00
    59.2257320         -0.0956200
     5.2139020          0.6414260
     2.0473460          0.4251530
S   3   1.00
     3.4471240          0.1964010
     0.4737850         -0.6923600
     0.1713210         -0.4261930
P   3   1.00
    64.0999580          0.1017890
    14.2871390          0.4561070
     3.8281350          0.6142820
P   3   1.00
     1.1039040          0.2359030
     0.1332360          0.3466000
     0.3991780          0.5580660
D   1   1.00
     0.6500000          1.0000000
****
```

Section S4. 6-31G* basis set file

```
-H     0
S   3   1.00
    18.7311370          0.03349460
     2.8253937          0.23472695
     0.6401217          0.81375733
S   1   1.00
     0.1612778          1.0000000
****
-B     0
S   6   1.00
  2068.8823000          0.0018663
   310.6495700          0.0142515
    70.6830330          0.0695516
    19.8610800          0.2325729
     6.2993048          0.4670787
     2.1270270          0.3634314
SP   3   1.00
     4.7279710         -0.1303938          0.0745976
     1.1903377         -0.1307889          0.3078467
     0.3594117          1.1309444          0.7434568
SP   1   1.00
     0.1267512          1.0000000          1.0000000
D   1   1.00
     0.6000000          1.0000000
****
-C     0
S   6   1.00
  3047.5249000          0.0018347
   457.3695100          0.0140373
   103.9486900          0.0688426
    29.2101550          0.2321844
     9.2866630          0.4679413
     3.1639270          0.3623120
SP   3   1.00
     7.8682724         -0.1193324          0.0689991
     1.8812885         -0.1608542          0.3164240
     0.5442493          1.1434564          0.7443083
SP   1   1.00
```

```
     0.1687144              1.0000000              1.0000000
D   1   1.00
     0.8000000              1.0000000
****
-N     0
S   6   1.00
   4173.5110000              0.0018348
    627.4579000              0.0139950
    142.9021000              0.0685870
     40.2343300              0.2322410
     12.8202100              0.4690700
      4.3904370              0.3604550
SP   3   1.00
     11.6263580             -0.1149610              0.0675800
      2.7162800             -0.1691180              0.3239070
      0.7722180              1.1458520              0.7408950
SP   1   1.00
      0.2120313              1.0000000              1.0000000
D   1   1.00
      0.8000000              1.0000000
****
-O     0
S   6   1.00
   5484.6717000              0.0018311
    825.2349500              0.0139501
    188.0469600              0.0684451
     52.9645000              0.2327143
     16.8975700              0.4701930
      5.7996353              0.3585209
SP   3   1.00
     15.5396160             -0.1107775              0.0708743
      3.5999336             -0.1480263              0.3397528
      1.0137618              1.1307670              0.7271586
SP   1   1.00
      0.2700058              1.0000000              1.0000000
D   1   1.00
      0.8000000              1.0000000
****
-F     0
S   6   1.00
   7001.7130900              0.0018196169
   1051.3660900              0.0139160796
    239.2856900              0.0684053245
     67.3974453              0.233185760
     21.5199573              0.471267439
      7.40310130             0.356618546
SP   3   1.00
     20.8479528             -0.108506975            0.0716287243
      4.80830834            -0.146451658            0.3459121030
      1.34406986             1.128688580            0.7224699570
SP   1   1.00
      0.358151393            1.0000000              1.0000000
D   1   1.00
      0.8000000              1.0000000
****
-Si     0
S   6   1.00
  16115.9000000              0.00195948
   2425.5800000              0.01492880
    553.8670000              0.07284780
    156.3400000              0.24613000
     50.0683000              0.48591400
     17.0178000              0.32500200
SP   6   1.00
```

```
    292.7180000             -0.00278094             0.00443826
     69.8731000             -0.03571460             0.03266790
     22.3363000             -0.11498500             0.13472100
      8.1503900             0.09356340             0.32867800
      3.1345800             0.60301700             0.44964000
      1.2254300             0.41895900             0.26137200
SP   3   1.00
      1.7273800             -0.24463000             -0.01779510
      0.5729220             0.00431572             0.25353900
      0.2221920             1.09818000             0.80066900
SP   1   1.00
      0.0778369             1.00000000             1.00000000
D   1   1.00
      0.4500000             1.0000000
****
-P     0
S   6   1.00
  19413.3000000             0.0018516
   2909.4200000             0.0142062
    661.3640000             0.0699995
    185.7590000             0.2400790
     59.1943000             0.4847620
     20.0310000             0.3352000
SP   6   1.00
    339.4780000             -0.00278217             0.00456462
     81.0101000             -0.0360499             0.03369360
     25.8780000             -0.1166310             0.13975500
      9.4522100             0.0968328             0.33936200
      3.6656600             0.6144180             0.45092100
      1.4674600             0.4037980             0.23858600
SP   3   1.00
      2.1562300             -0.2529230             -0.01776530
      0.7489970             0.0328517             0.27405800
      0.2831450             1.0812500             0.78542100
SP   1   1.00
      0.0998317             1.0000000             1.00000000
D   1   1.00
      0.5500000             1.0000000
****
-S     0
S   6   1.00
  21917.1000000             0.0018690
   3301.4900000             0.0142300
    754.1460000             0.0696960
    212.7110000             0.2384870
     67.9896000             0.4833070
     23.0515000             0.3380740
SP   6   1.00
    423.7350000             -0.0023767             0.0040610
    100.7100000             -0.0316930             0.0306810
     32.1599000             -0.1133170             0.1304520
     11.8079000             0.0560900             0.3272050
      4.6311000             0.5922550             0.4528510
      1.8702500             0.4550060             0.2560420
SP   3   1.00
      2.6158400             -0.2503740             -0.0145110
      0.9221670             0.0669570             0.3102630
      0.3412870             1.0545100             0.7544830
SP   1   1.00
      0.1171670             1.0000000             1.0000000
D   1   1.00
      0.6500000             1.0000000
****
-Cl     0
```

```
S   6   1.00
  25180.1000000             0.0018330
   3780.3500000             0.0140340
    860.4740000             0.0690970
    242.1450000             0.2374520
     77.3349000             0.4830340
     26.2470000             0.3398560
SP   6   1.00
    491.7650000            -0.0022974              0.0039894
    116.9840000            -0.0307140              0.0303180
     37.4153000            -0.1125280              0.1298800
     13.7834000             0.0450160              0.3279510
      5.4521500             0.5893530              0.4535270
      2.2258800             0.4652060              0.2521540
SP   3   1.00
      3.1864900            -0.2518300             -0.0142990
      1.1442700             0.0615890              0.3235720
      0.4203770             1.0601800              0.7435070
SP   1   1.00
      0.1426570             1.0000000              1.0000000
D   1   1.00
      0.7500000             1.0000000
****
```

Section S5. ACP file for HF-D3/MINIs

```
-H  0
H  1 0
l
7
2 0.120000 0.019752146768840
2 0.140000 -0.057681384248568
2 0.180000 0.077180835419727
2 0.240000 -0.048329611539726
2 0.400000 0.036004426723324
2 0.800000 -0.017954911921028
2 3.000000 -1.300152176570540
s
6
2 0.140000 0.053537123394618
2 0.220000 -0.154547284944162
2 0.280000 -0.012675123263846
2 0.600000 0.239049614006164
2 1.400000 0.536790738616504
2 3.000000 -0.240787240941970
-B  0
B  2 0
l
5
2 0.120000 0.008038455912135
2 0.200000 -0.039987145326567
2 0.400000 0.148500044059388
2 1.000000 -0.308374865579611
2 1.300000 -0.519836415312626
s
1
2 0.140000 0.089384845500027
p
2
2 0.120000 -0.020162020610331
2 0.200000 0.016596942666727
-C  0
C  2 0
```

```
l
8
2 0.120000 -0.012510454618930
2 0.140000 0.032454951132314
2 0.180000 -0.040283517341073
2 0.240000 0.049300798057028
2 0.400000 -0.156705879218398
2 0.500000 -0.069920266732228
2 0.900000 0.596533102491464
2 2.500000 -0.848732514272529
s
2
2 0.120000 0.050219021083560
2 0.200000 -0.175857157927506
p
6
2 0.120000 -0.008618787694833
2 0.200000 0.148016219818053
2 0.280000 0.070483864422903
2 0.600000 -0.058793967314518
2 1.100000 -0.399505952538211
2 2.500000 -0.521778345399917
-N  0
N  2 0
l
6
2 0.120000 0.004910332364775
2 0.140000 -0.006030200078750
2 0.300000 0.027668948238457
2 0.600000 -0.200733274751806
2 1.400000 0.315118979432174
2 1.500000 0.192855132256864
s
2
2 0.120000 -0.321691785175183
2 1.200000 0.045029305340688
p
6
2 0.120000 -0.005304154654419
2 0.160000 0.199832910951786
2 0.180000 0.109632184246528
2 0.500000 -0.045568939604577
2 0.700000 -0.395204060498005
2 0.900000 -0.324502993773902
-O  0
O  2 0
l
5
2 0.120000 0.006868100153429
2 0.160000 -0.028504598245850
2 0.260000 0.106305112544617
2 0.400000 -0.110567635240402
2 0.600000 -0.090929726560221
s
4
2 0.120000 0.192907605847906
2 0.200000 -0.261517489610578
2 0.500000 -0.460832996092277
2 1.900000 0.468815504636401
p
4
2 0.120000 -0.044226038798633
2 0.200000 0.024145940327329
2 0.240000 0.034383524268720
```

```
2 3.000000 -0.250957222296333
-F  0
F  2 0
l
7
2 0.120000 0.001201071696546
2 0.160000 -0.011094767690989
2 0.180000 -0.012601578298357
2 0.240000 0.076106442860530
2 0.400000 -0.035799638993456
2 0.700000 -0.253277836534745
2 1.100000 -0.243904108133977
s
2
2 0.160000 -0.298541417739927
2 1.300000 -0.559696251412562
p
2
2 0.120000 0.061009135865152
2 0.300000 0.065846281817406
-Si 0
Si 2 0
l
3
2 0.120000 0.024560309143780
2 0.180000 -0.072242108338752
2 0.400000 0.063487177778491
s
2
2 0.120000 -0.032653324458034
2 0.140000 -0.248418472884765
p
1
2 0.500000 0.059602779341734
-P  0
P  2 0
l
5
2 0.120000 0.052344656147227
2 0.140000 -0.096744207888736
2 0.200000 0.053355433675213
2 0.300000 -0.068146204264378
2 0.400000 -0.017139085339237
s
2
2 0.120000 0.215194763930423
2 0.600000 -1.261647266248893
p
2
2 0.140000 0.064047196792069
2 0.260000 -0.320418322687371
-S  0
S  2 0
l
5
2 0.120000 0.029858106050816
2 0.180000 -0.092902887048404
2 0.260000 0.040716255909233
2 0.500000 -0.251189152395120
2 0.600000 -0.158217377864688
s
1
2 0.120000 -0.260679569780629
p
```

```
2
2 0.120000 -0.069344807803546
2 0.200000 0.218717567740319
-Cl 0
Cl 2 0
l
4
2 0.120000 0.014803334648015
2 0.140000 -0.069311799637257
2 0.180000 0.161087877234957
2 0.300000 -0.307676734040921
s
2
2 0.220000 -0.139300352957365
2 0.260000 -0.415699199059726
p
2
2 0.120000 0.066904496312294
2 0.200000 -0.028207478717772
```

Section S6. ACP file for HF-D3/MINIX

```
-H  0
H  1 0
l
7
2 0.120000 0.014597530596173
2 0.140000 -0.041386316292299
2 0.180000 0.051333570911513
2 0.240000 -0.026507718774354
2 0.400000 0.031959213751350
2 0.700000 -0.034747459470645
2 3.000000 -1.134828627840867
s
7
2 0.140000 0.065682608206062
2 0.200000 -0.111639743815922
2 0.280000 -0.139629521792804
2 0.500000 0.215336198693111
2 0.700000 0.142989882161768
2 1.500000 0.478190293350417
2 3.000000 -0.249589814515530
-B  0
B  2 0
l
4
2 0.120000 0.004659899553511
2 0.180000 -0.001231004905479
2 0.300000 0.007404343664763
2 0.900000 -0.050322897597150
s
1
2 3.000000 0.355685249023271
p
3
2 0.120000 -0.007550721944642
2 0.400000 -0.062488507281968
2 0.500000 -0.009173354446546
-C  0
C  2 0
l
8
2 0.120000 -0.018199699341183
```

```
2 0.140000 0.072658759676635
2 0.160000 -0.080821858885667
2 0.220000 0.050475748122058
2 0.400000 -0.093782704373003
2 0.500000 -0.209091487392392
2 0.800000 0.562619846488431
2 1.800000 -0.740672373943856
s
2
2 0.120000 0.009972565474743
2 0.220000 -0.075828433519086
p
4
2 0.120000 0.008812229341457
2 0.220000 0.119099624013497
2 0.300000 0.090120895935993
2 0.800000 -0.290237952140879
-N  0
N  2 0
l
6
2 0.120000 0.003762406901759
2 0.140000 0.000433274182328
2 0.200000 -0.020733513443667
2 0.300000 0.056364789188455
2 0.600000 -0.129224876314160
2 0.700000 -0.176805210415714
s
2
2 0.120000 -0.389570228030583
2 3.000000 1.980650267019753
p
6
2 0.120000 -0.012471278004277
2 0.160000 0.296054202319582
2 0.180000 0.071414338013332
2 0.300000 -0.052023820068132
2 0.700000 -0.388122784886916
2 3.000000 0.415748544578656
-O  0
O  2 0
l
5
2 0.120000 0.005450943373969
2 0.160000 -0.022800597483110
2 0.240000 0.048337113373451
2 0.600000 -0.128054049804618
2 0.800000 -0.092746139042873
s
4
2 0.120000 0.120631249066454
2 0.180000 -0.041330745473994
2 0.400000 -0.573313655633566
2 1.500000 0.543632435136829
p
2
2 0.120000 -0.026632648636000
2 0.220000 0.058817157044142
-F  0
F  2 0
l
5
2 0.120000 -0.000198641853756
2 0.160000 -0.007562109185777
```

```
2 0.280000 0.042890504404431
2 0.600000 -0.222309573689598
2 1.100000 -0.630047795555416
s
1
2 0.140000 -0.310284174307246
p
4
2 0.120000 0.082387853977848
2 0.220000 0.120095683900399
2 0.600000 0.150118386237911
2 3.000000 0.177274854572769
-Si 0
Si 3 0
l
3
2 0.120000 0.019126759031882
2 0.180000 -0.029306576685440
2 0.400000 0.090144182754890
s
1
2 0.280000 -0.571792233365379
p
2
2 0.120000 0.110179371777462
2 0.240000 -0.161078339832869
d
2
2 0.120000 -0.075921447409180
2 1.800000 -0.000003138945012
-P  0
P  3 0
l
6
2 0.120000 0.038271043700104
2 0.140000 -0.119451574320205
2 0.200000 0.182616585624843
2 0.300000 -0.182394774837426
2 0.500000 0.241920491148716
2 0.800000 0.035922561800158
s
1
2 0.400000 -0.500804631732884
p
2
2 0.120000 0.047725729628080
2 0.220000 -0.307836142395298
d
3
2 0.120000 0.136503273399660
2 0.400000 -0.607110380326615
2 0.900000 -0.000016701044198
-S  0
S  3 0
l
4
2 0.140000 0.061108135583928
2 0.180000 -0.141328544570641
2 0.300000 0.195909466116472
2 1.000000 -0.534244353206738
s
3
2 0.120000 -0.004559625008032
2 0.300000 0.000002424919082
```

```
2 0.500000 -0.000012327252986
p
1
2 0.120000 -0.025183364220063
d
2
2 0.120000 -0.040498602097439
2 0.220000 -0.096654395991266
-Cl 0
Cl 3 0
l
8
2 0.120000 -0.005471671180392
2 0.160000 -0.000074916415328
2 0.180000 0.040444076745094
2 0.200000 0.003529111445177
2 0.400000 -0.257030192091142
2 0.500000 -0.000020251022917
2 0.600000 -0.005983968357407
2 1.400000 0.000014416188195
s
2
2 0.140000 -0.182102986095589
2 0.200000 -0.216935459596321
p
3
2 0.120000 0.044631937834784
2 0.220000 -0.000001867903720
2 0.500000 0.000025839813666
d
2
2 0.120000 0.000028067608271
2 0.140000 0.071078798634246
```

Section S7. ACP file for HF-3c

```
-H  0
H  1 0
l
8
2 0.120000 0.010912884073433
2 0.140000 -0.041139841974166
2 0.160000 0.034421379696394
2 0.300000 0.002129209726935
2 0.400000 0.016997502852076
2 0.600000 -0.022737866718080
2 1.200000 0.024697865044555
2 3.000000 -1.038149087364648
s
6
2 0.120000 0.067816574974078
2 0.180000 -0.124160009700526
2 0.240000 -0.054245296778459
2 0.800000 0.221319864592240
2 1.300000 0.560647180867972
2 3.000000 -0.419787644038363
-B  0
B  2 0
l
4
2 0.120000 0.005120962196134
2 0.160000 -0.001319293819505
2 0.300000 0.004577220751198
```

```
2 0.900000 -0.077533702976636
s
2
2 0.120000 -0.039004903056780
2 3.000000 1.740243752817217
p
2
2 0.120000 -0.000333672739675
2 0.500000 -0.038812870022048
-C  0
C  2 0
l
8
2 0.120000 -0.018766557577644
2 0.140000 0.071679764890841
2 0.160000 -0.079317131463539
2 0.220000 0.055791294065139
2 0.300000 -0.043867356082232
2 0.500000 -0.259934835370250
2 0.800000 0.650593359957896
2 2.000000 -0.681690661041336
s
2
2 0.120000 0.008155093881614
2 0.220000 -0.155260545580509
p
5
2 0.140000 0.035438443988530
2 0.200000 0.030156409274272
2 0.280000 0.137855015672249
2 0.700000 -0.035140232268219
2 1.000000 -0.451453100520033
-N  0
N  2 0
l
5
2 0.120000 0.001153625706010
2 0.140000 0.004289867195845
2 0.200000 -0.018749470258004
2 0.300000 0.031361710660123
2 0.600000 -0.093345174231203
s
2
2 0.120000 -0.424716239801033
2 3.000000 1.524915097217586
p
4
2 0.160000 0.311650294306681
2 0.200000 0.041832068486251
2 0.700000 -0.648460036263915
2 3.000000 0.471179015749668
-O  0
O  2 0
l
6
2 0.120000 0.006297936978142
2 0.140000 -0.016977983010809
2 0.200000 0.013166847213324
2 0.220000 0.009899770474453
2 0.500000 0.015734694878607
2 0.600000 0.045492005014255
s
4
2 0.120000 0.173545097875169
```

```
2 0.200000 -0.219627593408863
2 0.280000 -0.141436662377869
2 0.700000 -0.265077082912868
p
3
2 0.120000 -0.046932534582577
2 0.200000 0.104329953068143
2 0.600000 -0.279376666443716
-F  0
F  2 0
l
6
2 0.120000 -0.036978890694085
2 0.140000 0.065614593318295
2 0.180000 -0.049776466342839
2 0.300000 0.070583198584528
2 1.000000 -0.247464306956725
2 1.200000 -0.515003427225714
s
2
2 0.140000 -0.178239019626770
2 0.220000 -0.271771104474598
p
3
2 0.120000 0.127252026808109
2 0.160000 -0.021635992859766
2 0.500000 0.192283018373285
-Si 0
Si 3 0
l
3
2 0.120000 0.016036940817685
2 0.180000 -0.017559187650418
2 0.400000 0.052936359082537
s
1
2 0.280000 -0.544180748219028
p
2
2 0.120000 0.061533085067596
2 0.240000 -0.075428202912646
d
2
2 0.120000 -0.069853835432700
2 1.800000 -0.000002392884314
-P  0
P  3 0
l
5
2 0.120000 0.041063241668212
2 0.140000 -0.122330270714401
2 0.200000 0.155786751076008
2 0.300000 -0.089660572134811
2 0.600000 0.156075929283713
s
2
2 0.220000 -0.304117981085370
2 0.400000 -0.091689618588495
p
2
2 0.120000 0.007324657908489
2 0.220000 -0.236977232862017
d
4
```

```
2 0.120000 0.155759800340673
2 0.260000 -0.206657904283257
2 0.600000 -0.387269206938102
2 0.900000 -0.000016437956920
-S  0
S  3 0
l
5
2 0.120000 0.022726265364118
2 0.140000 0.014340300962228
2 0.180000 -0.101719065187035
2 0.300000 0.185684776450542
2 1.000000 -0.373360305514642
s
3
2 0.120000 -0.030784643641361
2 0.300000 0.000002704908452
2 0.500000 -0.000014089429351
p
1
2 0.120000 -0.057696215181526
d
1
2 0.120000 -0.099700367416487
-Cl 0
Cl 3 0
l
7
2 0.120000 0.000993752160937
2 0.160000 -0.000078651148566
2 0.180000 0.020972140822369
2 0.400000 -0.056563379773554
2 0.500000 -0.000020300234564
2 0.600000 -0.153451616948867
2 1.400000 0.000014895141836
s
2
2 0.120000 -0.099705081964320
2 0.140000 -0.233923768139323
p
3
2 0.120000 0.009796463144481
2 0.220000 -0.000001046963499
2 0.500000 0.000025671322322
d
2
2 0.120000 0.000027279924294
2 0.140000 0.045376983148558
```

Section S8. ACP file for HF-D3/6-31G*

```
-H  0
H  1 0
l
9
2 0.120000 0.005335263824203
2 0.140000 0.002289151082238
2 0.160000 -0.024442624349587
2 0.240000 0.057343154759566
2 0.400000 -0.163582609504628
2 0.600000 0.232406389876045
2 1.000000 -0.117991681174980
2 1.700000 0.196755579547757
```

```
2 3.000000 -0.971930661890001
s
6
2 0.120000 -0.099419693134682
2 0.140000 0.159713009134956
2 0.260000 -0.135121711525785
2 0.700000 0.232275003512689
2 1.400000 0.233181380119291
2 3.000000 -0.306924428872332
-B  0
B  3 0
l
4
2 0.120000 0.008114321089119
2 0.140000 -0.016315367444074
2 0.300000 0.015953032935283
2 0.500000 0.014999858525534
s
1
2 0.160000 -0.004579631692462
p
2
2 0.120000 0.023424892719621
2 0.220000 -0.026669999814328
d
2
2 0.120000 0.007652362681395
2 0.180000 -0.036950891757432
-C  0
C  3 0
l
9
2 0.120000 -0.010413643521740
2 0.140000 0.027617519896562
2 0.200000 -0.085691776671693
2 0.260000 0.056474927406062
2 0.280000 0.116107402624161
2 0.500000 -0.402084828590799
2 0.800000 0.348083790763103
2 0.900000 0.406388747205717
2 1.900000 -0.601264059614186
s
4
2 0.120000 -0.083151279939434
2 0.260000 0.024461923166834
2 0.300000 0.340214628144242
2 1.200000 -0.602265496569737
p
2
2 0.120000 0.013466088296609
2 0.280000 -0.057509300668460
d
3
2 0.120000 0.023980032708584
2 0.240000 -0.101492245116835
2 0.500000 0.126371142858247
-N  0
N  3 0
l
6
2 0.120000 -0.006022771612744
2 0.160000 0.015662799350099
2 0.220000 -0.024410249865171
2 0.300000 0.068270951249918
```

```
2 0.400000 0.011096842973182
2 0.600000 -0.259463960943656
s
2
2 0.120000 -0.020014537347287
2 0.500000 0.230418949635545
p
4
2 0.120000 -0.072539462315238
2 0.200000 0.156439717893717
2 0.220000 0.004525431815165
2 0.400000 0.071480675007927
d
3
2 0.120000 0.096472152845025
2 0.260000 -0.590499713137727
2 0.700000 1.990456300941629
-O  0
O  3 0
l
6
2 0.120000 0.006499066055270
2 0.140000 -0.017064168652508
2 0.200000 0.022511176844702
2 0.300000 0.016769938211214
2 0.700000 -0.156556467156598
2 3.000000 -0.258437931577082
s
4
2 0.120000 -0.014763172809516
2 0.300000 0.019492394631987
2 0.400000 0.356604053762252
2 1.800000 -1.025110802363867
p
4
2 0.120000 -0.010603212988619
2 0.180000 0.065757680555392
2 0.400000 -0.049995938674442
2 1.600000 0.245947331342133
d
5
2 0.120000 0.049455342983780
2 0.240000 -0.143426645866993
2 0.280000 -0.298871580930488
2 0.700000 1.328220068437189
2 1.100000 0.974585399708539
-F  0
F  3 0
l
5
2 0.120000 -0.007083187484638
2 0.140000 0.019778428077657
2 0.200000 -0.031288805461581
2 0.400000 0.076456857706809
2 0.800000 -0.269140374947919
s
2
2 0.140000 -0.220185691964973
2 0.160000 -0.028760705984404
p
3
2 0.120000 0.082689795824632
2 0.500000 -0.034302632669946
2 3.000000 0.262371356724822
```

```
d
1
2 0.260000 0.085174095266719
-Si 0
Si 3 0
l
3
2 0.140000 0.019070072967915
2 0.200000 -0.068669517066984
2 0.400000 0.043240811479710
s
0
p
2
2 0.120000 0.028818268580216
2 0.140000 0.019220450797341
d
2
2 0.120000 0.013269815705283
2 0.140000 0.001200671487240
-P  0
P  3 0
l
5
2 0.120000 -0.015749466885213
2 0.140000 -0.023676708017776
2 0.180000 0.092578405941051
2 0.300000 -0.009964066456330
2 0.600000 0.005220556945343
s
2
2 0.120000 0.129731302123130
2 0.260000 -0.327384006484185
p
2
2 0.120000 0.058074903532203
2 0.180000 -0.171016924197012
d
3
2 0.120000 0.148039655456661
2 0.200000 -0.281768961714683
2 0.240000 -0.084416214165550
-S  0
S  3 0
l
3
2 0.120000 0.003204463729525
2 0.180000 -0.021162135245606
2 0.300000 0.087659429040425
s
2
2 0.120000 -0.239089382041029
2 0.400000 0.167258759313735
p
2
2 0.120000 0.012672501735699
2 0.900000 -0.422042274324779
d
2
2 0.120000 0.125455671790176
2 0.180000 -0.272933663148112
-Cl 0
Cl 3 0
l
```

```
5
2 0.120000 -0.001936192424147
2 0.160000 0.004845372246876
2 0.220000 0.002341633069431
2 0.240000 0.032791435700485
2 1.100000 -0.220459694917153
s
2
2 0.120000 -0.003487708642873
2 0.140000 -0.323050814224878
p
2
2 0.120000 0.007320534420904
2 1.100000 -0.139692991182561
d
2
2 0.120000 0.078409873317561
2 0.220000 -0.172347009609912
```

Section S9. Formulas for all the statistical error measures

a) Mean absolute error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where, $x_i = \left|x_{calc,i} - x_{ref,i}\right|$

b) Mean signed error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where, $x_i = x_{calc,i} - x_{ref,i}$

c) Maximum absolute error (MAXE)

$$MAXE = \max_{i} \left|x_{calc,i} - x_{ref,i}\right|$$

d) Root-mean-square error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

where, $x_i = x_{calc,i} - x_{ref,i}$

e) Standard deviation (SD)

$$SD = \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

where, $x_i = x_{calc,i} - x_{ref,i}$

$$\overline{x} = \frac{1}{n}(x_{calc,i} - x_{ref,i})$$

Section S10. Extrapolation scheme used for generation of new reference data

We carried out the generation of new reference energies for some datasets used in this work at the DLPNO-CCSD(T)/CBS level of theory. The extrapolation scheme we used is a triple-ζ and quadruple-ζ extrapolation using def2-TZVPP and def2-QZVPP basis sets from Ahlrichs' family of basis sets. The extrapolation formulas chosen were similar to the one reported extensively in literature. At first a complete basis set (CBS) limit energy value is obtained at the HF level. Note that the HF/CBS extrapolation lacks the treatment of electronic correlation completely. Therefore, to correct for this shortcoming a CBS limit correlation energy value obtained at the RI-MP2 level is added to the energy value obtained with HF/CBS extrapolation. Finally, to correct for the remaining insufficient description of the electronic correlation, a difference in the correlation energy between CCSD(T) (here DLPNO-CCSD(T) for efficiency) and RI-MP2 using a triple-ζ basis set is added to the sums of energy obtained with HF/CBS extrapolation and correlation energy obtained with RI-MP2/CBS extrapolation. The extrapolation scheme just described is represented by the following formulas:

$$E = E_{HF/CBS} + E_{RI-MP2/CBS} + \delta\mathrm{CCSD(T)}$$

where,

$$E_{HF/CBS} = \frac{E_{HF/def2-TZVPP} * \exp\left(-\alpha\sqrt{Y}\right) - E_{HF/def2-QZVPP} * \exp\left(-\alpha\sqrt{X}\right)}{\exp\left(-\alpha\sqrt{Y}\right) - \exp\left(-\alpha\sqrt{X}\right)}$$

$$E_{RI-MP2/CBS} = \frac{E_{RI-MP2/def2-TZVPP} * X^{\beta} - E_{RI-MP2/def2-QZVPP}\, Y^{\beta}}{X^{\beta} - Y^{\beta}}$$

$$\delta\mathrm{CCSD(T)} = E_{DLPNO-CCSD(T)/def2-TZVPP} - E_{RI-MP2/def2-TZVPP}$$

The X and Y in the above equations refer to the cardinal number of the chosen basis set family i.e X=3 for the triple-ζ def2-TZVPP basis set and Y=4 for the quadruple-ζ def2-QZVPP basis set. The α and β parameters as obtained from literature were 7.88 and 2.97 for the Ahlrichs' family of basis sets.

For non-covalent interaction energies, the mean absolute error (MAE) of the DLPNO-CCSD(T)/CBS extrapolation scheme described above relative to conventional CCSD(T)/CBS for datasets such as A24 is 0.04 kcal/mol and for S22 is 0.07 kcal/mol. For molecular conformational energies, the MAE of the chosen scheme is only 0.18 kcal/mol relative to CCSD(T)/CBS reference data of the TPCONF dataset. For molecular deformation energies, the MAE of the chosen DLPNO-CCSD(T)/CBS scheme is only 0.02 kcal/mol for the MOLdef-H2O dataset (calculated in this work along with the Water-2body dataset using CCSD(T)/CBS technique similar to that in the Water38 dataset by Temelso *et al.*).

Table S1. Detailed list of data sets in the ACP training set.

Category	Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
Non-covalent interaction energies of molecular complexes:					
π-stacking	Pisub	105	Interaction energies of non-covalently bound substituted aromatic dimer complexes	DLPNO-CCSD(T)/CBS	123, 124, This work
	Pi29n	29	Interaction energies of neutral π-conjugated dimer complexes representing organic electronic precursors	CCSD(T)/CBS	125
	BzDC215	170	Interaction energies of dimer complexes (excluding neon and argon containing systems) of benzene with small molecules	CCSD(T)/CBS	126
	C2H4NT	75	Interaction energies of dimer complexes of ethene and coronene	CCSD(T)/CBS	127
Hydrophobic	ADIM6	6	Interaction energies of six alkane dimer complexes ranging from ethane to n-heptane	W1-F12	38, 128, 129
	HC12	12	Interaction energies of saturated and unsaturated hydrocarbon dimer complexes	CCSD(T)/CBS	130
Pnicogen-bonding	PNICO23	23	Interaction energies of dimer complexes representing pnictogen-bonding	W1-F12, W2-F12	128, 131
Halogen-bonding	Hill18	18	Interaction energies of hydrogen-bonded and halogen-bonded dimer complexes	CCSD(T)/CBS	132
	X40x10	220	Interaction energies of dimer complexes (excluding iodine and bromine containing systems) representing halogen-bonding at various intermolecular distances	CCSD(T)/CBS	133
Hydrogen-bonding	HBC6	118	Interaction energies of doubly hydrogen-bonded dimer complexes at various intermolecular distances	CCSD(T)/CBS	134, 135
	MiriyalaHB104	104	Interaction energies of hydrogen-bonded dimer complexes	CCSD(T)/CBS	136, 137
	IonicHB	96	Interaction energies of charged (both positive and negative) hydrogen-bonded dimer complexes at various intermolecular distances	CCSD(T)/CBS	138

Category	Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
	HB375x10	3749	Interaction energies of neutral hydrogen-bonded dimer complexes at various intermolecular distances	CCSD(T)/CBS	139
	IHB100x10	350	Interaction energies of charged (both positive and negative) hydrogen-bonded dimer complexes at various intermolecular distances	CCSD(T)/CBS	139
	HB300SPXx10	1980	Interaction energies of neutral hydrogen-bonded dimer complexes in an extended chemical space (excluding iodine and bromine containing systems) at various intermolecular distances.	CCSD(T)/CBS	140
	CARBHB12	12	Interaction energies of dimer complexes between singlet carbene analogues and H_2O, NH_3, HCl	W2-F12	128
Mixed NCIs	S22x5	110	Interaction energies of small non-covalently bound dimer complexes at various intermolecular distances	CCSD(T)/CBS	135, 141, 142
	S66x8	528	Interaction energies of small non-covalently bound dimer complexes at various intermolecular distances	CCSD(T)/CBS	143–145
	S66a8	528	Interaction energies of small non-covalently bound dimer complexes at various intermolecular angular displacements	CCSD(T)/CBS	144
	A21x12	228	Interaction energies of small non-covalently bound dimer complexes (excluding argon containing systems) at various intermolecular distances	CCSD(T)/CBS	3, 146, 147
	NBC10ext	195	Interaction energies of non-covalently interacting dimer complexes at various intermolecular distances	CCSD(T)/CBS	127, 135, 148–150
	3B-69-DIM	207	Interaction energies of all relevant pairs of monomers from 3B-69-TRIM	CCSD(T)/CBS	151
	3B-69-TRIM	69	Interaction energies of trimer complexes of small organic molecules	CCSD(T)/CBS	151
	HW30	30	Interaction energies of dimer complexes of hydrocarbons and water	CCSD(T)/CBS	152
Anionic	SSI-anionic, WatAA-anionic, HSG-anionic, PLF547-anionic, IonicHB-anionic, IHB100x10-anionic	575, 64, 4, 155, 24, 650	Interaction energies of only anion-neutral and anion-cation dimer complexes from earlier described datasets	Various	135, 138, 139, 153–156
	Ionic43-anionic	37	Interaction energies of anion-neutral and anion-cation dimer complexes (excluding sodium, potassium, and lithium containing complexes)	CCSD(T)/CBS	157

Category	Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
Biomolecule-Biomolecule	BBI	100	Interaction energies of peptide backbone-backbone dimer complexes	DW-CCSD(T)-F12/aug-cc-pV(D+d)z	153
	SSI	2805	Interaction energies of peptide sidechain-sidechain dimer complexes	DW-CCSD(T)-F12/aug-cc-pV(D+d)z	153
	NucTAA	454	Interaction energies of dimer complexes of amino acid and nucleotide	DLPNO-CCSD(T)/CBS	158–161, This work
	CarbhydBz	34	Interaction energies of carbohydrate-benzene dimer complexes	CCSD(T)/CBS	162
	CarbhydNaph	46	Interaction energies of carbohydrate-naphthalene dimer complexes	CCSD(T)/CBS	163
	CarbhydAroAA	48	Interaction energies of dimer complexes representing carbohydrate and aromatic amino acids	DLPNO-CCSD(T)/CBS	164, This work
	CarbhydAro	161	Interaction energies of dimer complexes representing carbohydrate and substituted aromatic molecule	DLPNO-CCSD(T)/CBS	165, This work
	WatAA	259	Interaction energies of dimer complexes representing interactions between water and amino acids	DLPNO-CCSD(T)/CBS	154, This work
	HSG	17	Interaction energies of dimer complexes representing protein-ligand interactions	CCSD(T)/CBS	135, 155
	PLF547	392	Interaction energies of dimer complexes representing protein-ligand interactions	DLPNO-CCSD(T)/CBS	156
	JSCH	124	Interaction energies of dimer complexes of nucleotide base pairs	CCSD(T)/CBS	141
	DNAstack	40	Interaction energies of stacked DNA base-pair steps	CBS(T)-F12-CP	166
	DNA2body	10	Interaction energies of nucleobase pairs	CBS(T)-F12-CP	166
	ACHC	54	Interaction energies of nucleobase stacking configurations	DW-CCSD(T**)-F12/aug-cc-pVDZ	167
	BDNA	71	Interaction energies of nucleobase stacking configurations	CCSD(T)/CBS	168
	NucBTrimer	141	Interaction energies of complexes of nucleobase trimers	DLPNO-CCSD(T)/CBS	169, This work
Gas-Ligand	CH4PAH	382	Interaction energies of dimer complexes of methane and polycyclic aromatic hydrocarbons	CCSD(T)/CBS	170, 171
	CO2MOF	20	Interaction energies of dimer complexes of carbon dioxide and organic building units of metal-organic frameworks	inc-CCSD(T)|MP2+F12+INT/cc-pVDZ-F12	172
	CO2PAH	249	Interaction energies of dimer complexes of carbon dioxide and polycyclic aromatic hydrocarbons	CCSD(T**)-F12avg/CBS	173

Category	Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
	CO2NPHAC	96	Interaction energies of dimer complexes of carbon dioxide and nitrogen-doped poly-heterocyclic aromatic compounds	CCSD(T)/CBS	174
	BzGas	129	Interaction energies of nine benzene-gas dimer complexes at various intermolecular distances (where gas = CO_2, CH_4, N_2)	CCSD(T)/CBS	175
Water-Water	Water38	38	Interaction energies of water clusters $(H_2O)_n$ (where n = 2-10)	CCSD(T)/CBS	176
	Water1888	1888	Interaction energies of various water dimer configurations with reference data lying between -5 to +5 kcal/mol	CCSD(T)/CBS	127,177–179
	Water-2body	410	Interaction energies of various water dimer configurations	CCSD(T)/CBS	67, This work
BFSiPSCl	B-set	160	Interaction energies of dimer complexes containing boron at various intermolecular distances	DLPNO-CCSD(T)/CBS	64, This work
	F-set	160	Interaction energies of dimer complexes containing fluorine at various intermolecular distances	DLPNO-CCSD(T)/CBS	64, This work
	Si-set	152	Interaction energies of dimer complexes containing silicon at various intermolecular distances	DLPNO-CCSD(T)/CBS	64, This work
	P-set	120	Interaction energies of dimer complexes containing phosphorus at various intermolecular distances	DLPNO-CCSD(T)/CBS	64, This work
	S-set	144	Interaction energies of dimer complexes containing sulfur at various intermolecular distances	DLPNO-CCSD(T)/CBS	64, This work
	Cl-set	160	Interaction energies of dimer complexes containing chlorine at various intermolecular distances	DLPNO-CCSD(T)/CBS	64, This work
	Sulfurx8	104	Interaction energies of dimer complexes containing divalent sulfur at various intermolecular distances	CCSD(T)/CBS	180
Molecular conformational energies:					
Small molecule	37Conf8	258	Relative energies of conformers of organic molecule isomers	DLPNO-CCSD(T)/cc-pVTZ	181
	DCONF	2142	Relative energies of conformers of 62 model systems representing drug-like molecules at various intramolecular torsion angles	CCSD(T)/CBS	182
	ICONF	17	Relative energies of conformers of 10 molecules containing H, N, O, Si, P, and S	W1-F12	128
	MCONF	51	Relative energies of conformers of melatonin	CCSD(T)/CBS	183
	Torsion21	189	Relative energies of conformers of Glyoxal, Oxalyl halides, and their thiocarbonyl derivatives (excluding bromine containing systems) at	CCSD(T)/CBS	184

Category	Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
			various intramolecular torsion angles		
	MolCONF	5623	Relative energies of conformers of molecules taken from crystal structure database and protein-ligand database (only containing our 10 target elements)	DLPNO-CCSD(T)/cc-pVTZ	185
	ANI1ccxCONF	32944	Relative energies with respect to the energy minimum of organic molecules generated using normal mode sampling, dimer sampling, and torsion sampling	CCSD(T)*/CBS	186
Negatively charged	PEPCONF-Dipeptide-anionic, MolCONF-anionic	175, 79	Relative energies of conformers of systems containing only negative charge from earlier described datasets	Various	185, 187, This work
Biomolecule	PEPCONF-Dipeptide	875	Relative energies of conformers of various model dipeptide systems	DLPNO-CCSD(T)/CBS	187, This work
	TPCONF	8	Relative energies of conformers of two model tetrapeptides	CCSD(T)/CBS	188
	P76	71	Relative energies of conformers of five isolated small peptides containing aromatic side chains	CCSD(T)/CBS	189
	YMPJ	495	Relative energies of conformers of proteinogenic amino acid monomers	MP2-F12/cc-pVTZ-F12+[CCSD(Ts)-F12b – MP2-F12]/cc-pVDZ-F12	190
	SPS	17	Relative energies of conformers of DNA sugar-phosphate-sugar backbone	CCSD(T)/CBS	191
	rSPS	45	Relative energies of conformers of RNA sugar-phosphate-sugar backbone	CCSD(T)/CBS	192
	UpU46	45	Relative energies of conformers of model RNA backbone	DLPNO-CCSD(T)/CBS*	193
	SCONF	17	Relative energies of conformers of two model carbohydrates	CCSD(T)/CBS	128, 194
	DSCONF	27	Relative energies of conformers of three disaccharides	CCSD(T)/CBS	195
	SacchCONF	56	Relative energies of conformers of monosaccharides	CCSD(T)/CBS	196
	CCONF	426	Relative energies of conformers of glucose and α-maltose isomers	DLPNO-CCSD(T)/CBS	197
Hydrocarbon	ACONF	15	Relative energies of conformers of n-alkane chains	W1h-val	198
	BCONF	64	Relative energies of conformers of butane-1,4-diol	CCSD(T)-F12b/cc-pVTZ-F12	199
	PentCONF	342	Relative energies of conformers of n-pentane	CCSD(T)-F12/CBS	200
$(H_2O)_{11}$	Undecamer125	124	Relative energies of conformers of $(H_2O)_{11}$	CCSD(T)/CBS	201
Molecular deformation energies:					
Deformation	MOLdef	9298	Molecular deformation energies relative to the equilibrium geometry	DLPNO-CCSD(T)/CBS	64, This work

Category	Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
			of systems containing our 10 target elements		
	MOLdef-H2O	990	Molecular deformation energies relative to the equilibrium geometry of water containing systems	CCSD(T)/CBS	202, 203, This work

Table S2. Detailed list of data sets in the ACP validation set.

Category	Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
Non-covalent interaction energies:					
Mixed NCIs	BlindNCI	80	Interaction energies of 10 dimer complexes at various intermolecular distances used previously for blind test	CCSD(T)/CBS	204
	DES15K	11474	Interaction energies of various non-covalently bound dimer complexes	CCSD(T)/CBS	205
	NENCI-2021	5859	Interaction energies of non-equilibrium dimer complexes	CCSD(T)/CBS	206
Hydrogen-bonding	CE20	20	Interaction energies of water, ammonia, and hydrogen fluoride clusters	W1-F12	207, 208
	WaterOrg	2376	Interaction energies of hydrogen-bonding interactions between water clusters and organic molecule complexes	DLPNO-CCSD(T)/CBS	209
Halogen-bonding	XB45	33	Interaction energies of halogen-bonded dimer complexes	CCSD(T)/aug-cc-pVTZ	210
Chalcogen-bonding	CHAL336	48	Interaction energies of chalcogen-bonded dimer complexes	W1-F12 or DLPNO-CCSD(T)/CBS	211
Repulsive contacts	R160x6	960	Interaction energies of small dimer complexes at short intermolecular distances	CCSD(T)/CBS	212
	R739x5	4330	Interaction energies of small dimer complexes at short intermolecular distances	CCSD(T)/CBS	213
Anionic	HW6Cl-anionic	6	Interaction energies of clusters of $Cl^-(H_2O)_n$ (where n = 1-6)	CCSD(T)/CBS	214, 215
	HW6F-anionic	6	Interaction energies of clusters of $F^-(H_2O)_n$ (where n = 1-6)	CCSD(T)/CBS	214, 215
	FmH2O10-anionic	10	Interaction energies of clusters of $F^-(H_2O)_{10}$	CCSD(T)/CBS	214, 215
	SW49Bind345-anionic	30	Interaction energies of clusters of $SO_4^{2-}(H_2O)_n$ (where n = 3-5)	CCSD(T)/CBS	216
	SW49Bind6-anionic	18	Interaction energies of clusters of $SO_4^{2-}(H_2O)_6$	CCSD(T)/CBS	216
	Anionpi-anionic	16	Interaction energies of anion-π type non-covalently interacting dimer complexes	DLPNO-CCSD(T)/CBS	217
	IL236-anionic	236	Interaction energies of ion pair dimer complexes representing model ionic liquids	CCSD(T)/CBS	218

Category	**Data set(s)**	**Data points**	**Data set description**	**Reference energy level**	**Reference # in article**
	DES15K-anionic, NENCI-2021-anionic, CHAL336-anionic, XB45-anionic, S30L-anionic	1281, 889, 19, 12, 2	Interaction energies of only anion-neutral and anion-cation dimer complexes from earlier described datasets	Various	205, 206, 210, 211, 219
$(H_2O)_{20}$ cluster	H2O20Bind10	10	Interaction energies of clusters of $(H_2O)_{20}$	CCSD(T)/CBS	215
C_{60} dimer	C60dimer	14	Interaction energies of the C_{60} dimer complex at various intermolecular distances	DLPNO-CCSD(T)/CBS	220
Large molecule	L7	7	Interaction energies of seven relatively large non-covalently bound complexes	DLPNO-CCSD(T)/CBS	221, 222
	S12L	10	Interaction energies of supramolecular host-guest complexes	DLPNO-CCSD(T)/CBS	9, 11, 222
	S30L	26	Interaction energies of supramolecular host-guest complexes	Experimental back-corrected	219
	Ni2021	11	Interaction energies of large non-covalently bound complexes ranging in size between 126-1027 atoms	CIM-DLPNO-CCSD(T)‖RI-MP2	223
Molecular conformational energies:					
Small molecule	SafroleCONF	5	Relative energies of safrole conformers	CCSD(T)/CBS	224
	AlcoholCONF	31	Relative energies of small alcohol conformers	CCSD(T)/aug-cc-pVTZ	225
	BeranCONF	50	Relative energies of flexible organic molecule conformers relevant in crystal structure prediction	CCSD(T)/CBS or MP2D/CBS	226
	Torsion30	2107	Relative energies of conformers of model systems representing biaryl drug-like molecules at various intramolecular torsion angles (excluding many datapoints for which geometries were missing).	CCSD(T)*/CBS	227
Proteinogenic	MPCONF196	112	Relative energies of medium-sized macrocyclic peptide conformers	DLPNO-CCSD(T)/CBS	228
	PEPCONF-Tripeptide	647	Relative energies of conformers of various model tripeptide systems	DLPNO-CCSD(T)/CBS	187, This work
	PEPCONF-Disulfide	620	Relative energies of conformers of various model peptide systems containing disulfide linkages	LC-ωPBE-XDM/aug-cc-pVTZ	187
	PEPCONF-Cyclic	320	Relative energies of conformers of various macrocyclic peptides	LC-ωPBE-XDM/aug-cc-pVTZ	187
	PEPCONF-Bioactive	175	Relative energies of conformers of various polypeptides that show bioactive function	LC-ωPBE-XDM/aug-cc-pVTZ	187
Negatively charged	PEPCONF-Disulfide-anionic, PEPCONF-Bioactive-anionic	150, 20	Relative energies of conformers of systems containing only negative charge from earlier described datasets	LC-ωPBE-XDM/aug-cc-pVTZ	187

Table S3. Detailed error analysis with respect to reference data in the training set. The numbers in bracket in the first column indicates the number of data points. The various shorthand notations are as follows: MINIs = HF-D3/MINIs, MINIs-ACP = HF-D3/MINIs-ACP, MINIX = HF-D3/MINIX, MINIX-ACP = HF-D3/MINIX-ACP, 6-31G* = HF-D3/6-31G*, 6-31G*-ACP = HF-D3/6-31G*-ACP, MAE = mean absolute error in kcal/mol, MSE = mean signed error in kcal/mol, MAXE = maximum absolute error in kcal/mol, RMSE = root-mean-square error in kcal/mol, and SD = standard deviation in kcal/mol. The "fit" represents results from LASSO fitting and "scf" represents results of actual self-consistent field calculations.

Dataset (# of data points)		**MINIs**	**MINIs-ACP (fit)**	**MINIs-ACP (scf)**	**MINIX**	**MINIX-ACP (fit)**	**MINIX-ACP (scf)**	**HF-3c**	**HF-3c-ACP (fit)**	**HF3c-ACP (scf)**	**6-31G***	**6-31G*-ACP (fit)**	**6-31G*-ACP (scf)**
S22x5 (110)	MAE	1.40	0.35	0.36	1.40	0.35	0.36	0.53	0.38	0.40	1.56	0.27	0.26
	MSE	-1.18	0.18	0.11	-1.18	0.17	0.13	0.17	0.19	0.16	-1.55	-0.11	-0.12
	MAXE	8.26	3.05	2.90	8.26	3.07	2.94	2.62	3.08	2.95	5.40	1.25	1.02
	RMSE	2.42	0.57	0.57	2.42	0.57	0.58	0.80	0.62	0.63	2.21	0.40	0.36
	SD	2.12	0.55	0.57	2.12	0.55	0.57	0.79	0.59	0.61	1.57	0.38	0.35
S66x8 (528)	MAE	1.24	0.28	0.25	1.24	0.26	0.26	0.37	0.27	0.27	1.49	0.21	0.23
	MSE	-1.15	0.14	0.08	-1.15	0.12	0.09	0.08	0.11	0.07	-1.49	-0.02	-0.03
	MAXE	6.68	1.64	1.51	6.68	1.77	1.72	2.46	1.99	1.97	5.09	1.18	0.99
	RMSE	1.78	0.42	0.36	1.78	0.38	0.38	0.56	0.39	0.41	1.88	0.29	0.30
	SD	1.36	0.39	0.35	1.36	0.36	0.37	0.55	0.38	0.40	1.15	0.29	0.30
S66a8 (528)	MAE	1.18	0.25	0.24	1.18	0.25	0.25	0.36	0.25	0.26	1.64	0.20	0.23
	MSE	-1.18	0.08	0.01	-1.18	0.05	0.01	0.06	0.04	0.00	-1.64	-0.06	-0.07
	MAXE	3.93	3.04	2.87	3.93	2.86	2.78	4.07	2.95	2.86	5.19	0.75	0.76
	RMSE	1.37	0.41	0.37	1.37	0.38	0.38	0.54	0.39	0.40	1.91	0.25	0.27
	SD	0.70	0.41	0.37	0.70	0.38	0.39	0.54	0.39	0.40	0.98	0.24	0.26
A21x12 (228)	MAE	0.26	0.12	0.11	0.26	0.12	0.11	0.19	0.14	0.12	0.44	0.09	0.09
	MSE	-0.19	0.07	0.03	-0.19	0.07	0.04	0.12	0.09	0.05	-0.44	-0.03	-0.04
	MAXE	3.84	2.05	1.18	3.84	2.24	1.27	3.27	2.48	1.80	3.08	1.06	0.97
	RMSE	0.62	0.24	0.18	0.62	0.26	0.20	0.42	0.30	0.23	0.78	0.16	0.16
	SD	0.59	0.23	0.18	0.59	0.25	0.20	0.40	0.29	0.23	0.65	0.16	0.16
NBC10ext (195)	MAE	1.28	0.32	0.32	1.25	0.32	0.33	0.54	0.30	0.30	0.60	0.24	0.29
	MSE	-1.27	-0.23	-0.23	-1.24	-0.22	-0.23	-0.50	-0.17	-0.18	-0.58	-0.18	-0.27
	MAXE	5.05	1.50	1.44	5.05	1.37	1.40	3.30	1.90	1.93	2.40	0.80	0.82
	RMSE	1.83	0.49	0.49	1.80	0.49	0.49	0.90	0.45	0.46	0.83	0.30	0.36
	SD	1.31	0.44	0.44	1.30	0.43	0.44	0.75	0.42	0.43	0.59	0.24	0.24
Sulfurx8 (104)	MAE	0.50	0.39	0.41	0.36	0.29	0.29	0.71	0.34	0.35	0.75	0.19	0.19
	MSE	-0.40	0.27	0.28	-0.24	0.21	0.21	0.66	0.23	0.23	-0.74	0.08	0.06
	MAXE	2.42	1.73	1.86	2.01	1.14	1.11	5.13	1.39	1.39	2.31	0.72	0.74
	RMSE	0.75	0.53	0.56	0.57	0.38	0.38	1.28	0.44	0.45	0.96	0.26	0.26
	SD	0.64	0.45	0.49	0.52	0.31	0.32	1.10	0.37	0.38	0.62	0.24	0.26
3B-69-DIM (207)	MAE	1.08	0.46	0.42	1.08	0.45	0.41	0.50	0.45	0.42	1.44	0.25	0.25
	MSE	-1.01	0.32	0.22	-1.01	0.30	0.22	0.28	0.30	0.22	-1.43	0.11	0.11
	MAXE	6.12	4.21	3.79	6.12	4.33	3.91	3.18	4.43	4.01	4.41	1.17	1.20
	RMSE	1.68	0.75	0.65	1.68	0.74	0.64	0.72	0.75	0.65	1.80	0.34	0.34
	SD	1.34	0.68	0.62	1.34	0.67	0.60	0.67	0.69	0.62	1.10	0.32	0.33
3B-69-TRIM (69)	MAE	3.14	1.08	0.94	3.14	1.05	0.93	1.13	1.05	0.93	4.32	0.59	0.60
	MSE	-3.12	0.88	0.58	-3.12	0.83	0.59	0.75	0.83	0.59	-4.32	0.31	0.30
	MAXE	9.00	6.98	6.10	9.00	7.17	6.24	4.96	7.27	6.33	10.07	1.92	2.02
	RMSE	3.93	1.55	1.31	3.93	1.52	1.31	1.50	1.54	1.32	4.69	0.77	0.78
	SD	2.40	1.29	1.18	2.40	1.29	1.18	1.31	1.31	1.19	1.84	0.71	0.73
WatAA (259)	MAE	3.54	0.80	0.76	3.54	0.69	0.68	1.25	0.69	0.67	3.29	0.34	0.27

Dataset (# of data points)		MINIs	MINIs-ACP (fit)	MINIs-ACP (scf)	MINIX	MINIX-ACP (fit)	MINIX-ACP (scf)	HF-3c	HF-3c-ACP (fit)	HF3c-ACP (scf)	6-31G*	6-31G*-ACP (fit)	6-31G*-ACP (scf)
	MSE	-3.54	-0.06	-0.20	-3.54	-0.03	-0.09	-0.79	-0.08	-0.15	-3.29	-0.22	-0.11
	MAXE	8.08	2.71	2.87	8.08	2.61	2.70	5.35	2.75	2.83	7.52	1.14	1.06
	RMSE	3.90	1.02	1.02	3.90	0.91	0.92	1.80	0.93	0.94	3.41	0.45	0.36
	SD	1.64	1.02	1.00	1.63	0.91	0.92	1.62	0.92	0.93	0.91	0.39	0.35
BBI (100)	MAE	1.04	1.07	1.00	1.04	1.03	1.07	0.88	1.05	1.10	1.69	0.46	0.50
	MSE	-1.03	1.07	1.00	-1.03	1.03	1.07	0.88	1.05	1.10	-1.67	0.46	0.50
	MAXE	2.33	2.02	1.93	2.33	2.25	2.38	1.79	2.22	2.38	2.64	1.10	0.98
	RMSE	1.18	1.15	1.08	1.18	1.13	1.17	0.92	1.14	1.20	1.79	0.49	0.53
	SD	0.57	0.43	0.39	0.57	0.47	0.48	0.25	0.45	0.47	0.64	0.16	0.18
SSI (2805)	MAE	0.87	0.21	0.21	0.87	0.21	0.20	0.28	0.22	0.22	0.76	0.14	0.15
	MSE	-0.83	0.12	0.10	-0.82	0.13	0.11	-0.03	0.14	0.12	-0.75	0.03	-0.04
	MAXE	6.86	10.84	10.39	6.86	3.69	3.68	4.76	3.82	3.83	4.45	3.06	3.25
	RMSE	1.06	0.43	0.42	1.05	0.37	0.37	0.49	0.38	0.38	0.95	0.22	0.22
	SD	0.66	0.41	0.41	0.66	0.35	0.35	0.49	0.35	0.36	0.58	0.22	0.22
JSCH (124)	MAE	2.60	0.63	0.73	2.56	0.64	0.73	0.98	0.66	0.77	2.30	0.45	0.41
	MSE	-2.52	0.18	-0.03	-2.48	0.17	0.01	0.15	0.18	0.03	-2.24	0.22	0.21
	MAXE	13.82	3.36	4.19	13.82	3.57	4.01	6.12	3.62	4.03	9.05	2.20	1.94
	RMSE	3.80	0.87	1.03	3.76	0.90	1.04	1.36	0.92	1.08	3.13	0.63	0.57
	SD	2.86	0.86	1.04	2.84	0.89	1.05	1.36	0.90	1.08	2.19	0.59	0.53
DNAstack (40)	MAE	0.87	0.37	0.38	0.87	0.34	0.34	0.42	0.33	0.34	0.90	0.21	0.19
	MSE	-0.77	0.23	0.22	-0.77	0.23	0.22	0.37	0.22	0.21	-0.78	0.17	0.09
	MAXE	2.42	1.32	1.31	2.42	1.39	1.32	1.33	1.45	1.34	2.19	0.71	0.58
	RMSE	1.12	0.48	0.48	1.12	0.45	0.45	0.55	0.44	0.44	1.12	0.27	0.23
	SD	0.82	0.42	0.43	0.82	0.39	0.40	0.40	0.39	0.39	0.81	0.21	0.21
DNA2body (10)	MAE	4.48	0.44	0.48	4.48	0.45	0.50	0.21	0.48	0.55	3.63	0.27	0.33
	MSE	-4.48	-0.44	-0.48	-4.48	-0.45	-0.50	0.09	-0.48	-0.55	-3.63	0.04	-0.28
	MAXE	5.16	1.09	1.18	5.16	1.11	1.21	0.37	1.11	1.22	4.48	0.47	0.76
	RMSE	4.49	0.55	0.61	4.49	0.54	0.60	0.24	0.55	0.62	3.66	0.31	0.42
	SD	0.30	0.36	0.39	0.30	0.31	0.34	0.23	0.29	0.30	0.48	0.33	0.33
ACHC (54)	MAE	1.44	0.45	0.46	1.44	0.44	0.45	0.28	0.41	0.42	1.15	0.30	0.38
	MSE	-1.44	-0.43	-0.44	-1.44	-0.42	-0.43	0.06	-0.39	-0.40	-1.15	-0.29	-0.37
	MAXE	5.62	0.95	1.01	5.62	0.92	0.95	1.48	0.88	0.93	2.36	0.52	0.62
	RMSE	1.66	0.51	0.52	1.66	0.50	0.51	0.37	0.47	0.49	1.22	0.33	0.40
	SD	0.83	0.27	0.28	0.83	0.27	0.28	0.36	0.27	0.27	0.43	0.15	0.15
BDNA (71)	MAE	2.08	0.37	0.50	2.08	0.35	0.48	0.60	0.37	0.50	1.69	0.14	0.13
	MSE	-2.05	-0.03	-0.17	-2.05	-0.01	-0.15	-0.09	-0.02	-0.16	-1.65	0.12	0.09
	MAXE	7.81	0.95	1.53	7.81	0.93	1.36	2.22	0.97	1.45	4.57	0.43	0.35
	RMSE	3.11	0.41	0.65	3.11	0.40	0.61	0.81	0.42	0.64	2.30	0.18	0.16
	SD	2.36	0.42	0.63	2.36	0.40	0.60	0.81	0.42	0.63	1.62	0.13	0.13
NucBTrimer (141)	MAE	10.27	1.69	1.37	10.27	1.60	1.49	1.73	1.60	1.56	10.60	1.02	1.06
	MSE	-10.27	1.13	0.40	-10.27	1.02	0.68	0.17	0.99	0.73	-10.60	0.56	0.70
	MAXE	19.17	7.21	6.58	19.17	7.21	7.59	6.79	7.42	8.04	17.12	2.83	2.64
	RMSE	10.71	1.95	1.79	10.71	1.90	1.89	2.35	1.90	1.95	10.87	1.18	1.24
	SD	3.05	1.59	1.75	3.05	1.61	1.77	2.35	1.63	1.81	2.42	1.04	1.02
NucTAA (454)	MAE	1.29	0.92	0.90	1.22	0.86	0.84	0.92	0.87	0.85	1.44	0.41	0.38
	MSE	-0.80	0.68	0.62	-0.75	0.67	0.64	0.62	0.68	0.65	-1.38	0.15	0.08
	MAXE	18.64	9.26	9.55	18.78	7.05	6.96	12.56	7.01	6.95	7.49	3.54	3.26
	RMSE	2.26	1.42	1.40	2.13	1.26	1.24	1.54	1.28	1.26	1.89	0.64	0.60
	SD	2.12	1.25	1.25	1.99	1.07	1.06	1.42	1.08	1.08	1.30	0.62	0.59
CarbhydBz (34)	MAE	2.96	0.49	0.56	2.96	0.50	0.54	0.62	0.52	0.56	2.53	0.24	0.18
	MSE	-2.96	-0.33	-0.42	-2.96	-0.35	-0.41	-0.54	-0.40	-0.44	-2.53	0.24	-0.06
	MAXE	3.69	1.24	1.18	3.69	1.18	1.12	0.95	1.11	1.07	3.36	1.27	1.02

Dataset (# of data points)		MINIs	MINIs-ACP (fit)	MINIs-ACP (scf)	MINIX	MINIX-ACP (fit)	MINIX-ACP (scf)	HF-3c	HF-3c-ACP (fit)	HF3c-ACP (scf)	6-31G*	6-31G*-ACP (fit)	6-31G*-ACP (scf)
	RMSE	3.01	0.54	0.60	3.01	0.54	0.59	0.67	0.57	0.60	2.58	0.32	0.25
	SD	0.54	0.43	0.44	0.54	0.42	0.43	0.39	0.41	0.41	0.49	0.22	0.24
CarbhydNaph (46)	MAE	3.72	0.54	0.66	3.72	0.58	0.66	0.61	0.66	0.73	2.97	0.42	0.12
	MSE	-3.72	-0.54	-0.65	-3.72	-0.57	-0.65	-0.61	-0.66	-0.73	-2.97	0.42	0.04
	MAXE	4.83	0.95	1.05	4.83	1.14	1.22	0.94	1.24	1.31	4.26	0.74	0.42
	RMSE	3.77	0.59	0.70	3.77	0.62	0.70	0.64	0.71	0.77	3.05	0.44	0.15
	SD	0.66	0.24	0.26	0.66	0.25	0.26	0.19	0.25	0.26	0.69	0.14	0.15
CarbhydAroAA (48)	MAE	1.61	0.26	0.28	1.61	0.28	0.29	0.35	0.28	0.30	2.55	0.71	0.85
	MSE	-1.60	-0.13	-0.17	-1.61	-0.15	-0.18	-0.25	-0.15	-0.18	-2.55	-0.71	-0.85
	MAXE	3.64	0.83	0.85	3.64	1.02	1.04	0.69	1.00	1.02	5.15	1.47	1.70
	RMSE	1.81	0.33	0.35	1.82	0.35	0.36	0.40	0.36	0.37	2.78	0.79	0.94
	SD	0.85	0.30	0.31	0.86	0.31	0.32	0.31	0.33	0.33	1.12	0.35	0.41
CarbhydAro (161)	MAE	4.05	0.44	0.52	4.05	0.41	0.46	0.56	0.47	0.53	4.16	0.18	0.26
	MSE	-4.05	-0.36	-0.48	-4.05	-0.32	-0.41	-0.52	-0.42	-0.49	-4.16	0.01	-0.20
	MAXE	7.25	1.23	1.36	7.25	1.16	1.29	1.34	1.47	1.66	7.42	0.79	0.82
	RMSE	4.23	0.51	0.61	4.23	0.48	0.54	0.64	0.56	0.62	4.35	0.23	0.32
	SD	1.22	0.37	0.38	1.22	0.36	0.36	0.39	0.36	0.37	1.28	0.23	0.25
HSG (17)	MAE	0.94	0.19	0.18	0.94	0.19	0.19	0.33	0.19	0.19	0.89	0.12	0.12
	MSE	-0.86	0.11	0.09	-0.86	0.10	0.09	0.08	0.12	0.12	-0.89	0.09	0.01
	MAXE	1.72	0.56	0.47	1.72	0.55	0.51	1.36	0.57	0.56	2.64	0.40	0.39
	RMSE	1.01	0.23	0.21	1.01	0.22	0.22	0.47	0.23	0.24	1.04	0.16	0.16
	SD	0.56	0.20	0.20	0.56	0.21	0.21	0.48	0.21	0.21	0.56	0.14	0.16
PLF547 (392)	MAE	0.82	0.43	0.42	0.78	0.42	0.40	0.47	0.43	0.42	0.94	0.21	0.20
	MSE	-0.59	0.28	0.24	-0.56	0.29	0.26	0.30	0.31	0.28	-0.87	0.09	0.04
	MAXE	6.61	3.88	3.48	6.14	3.88	3.61	5.91	4.19	3.94	6.54	2.43	1.92
	RMSE	1.33	0.70	0.67	1.27	0.69	0.65	0.82	0.71	0.67	1.40	0.35	0.33
	SD	1.19	0.65	0.62	1.14	0.63	0.60	0.77	0.63	0.61	1.10	0.34	0.32
HBC6 (118)	MAE	3.27	0.62	0.53	3.27	0.63	0.54	1.13	0.65	0.60	2.79	0.37	0.30
	MSE	-2.84	0.41	0.17	-2.84	0.30	0.14	-0.01	0.26	0.07	-2.79	-0.16	-0.05
	MAXE	11.34	1.79	1.72	11.34	2.20	1.82	4.75	2.28	2.07	5.43	1.72	1.66
	RMSE	4.65	0.78	0.67	4.65	0.81	0.70	1.48	0.83	0.77	3.33	0.48	0.42
	SD	3.70	0.67	0.65	3.70	0.76	0.69	1.48	0.79	0.77	1.82	0.46	0.42
MiriyalaHB104 (104)	MAE	2.41	0.44	0.45	2.41	0.43	0.47	0.52	0.44	0.52	2.44	0.20	0.19
	MSE	-2.41	-0.10	-0.26	-2.41	-0.09	-0.22	-0.28	-0.17	-0.31	-2.44	-0.01	0.04
	MAXE	5.63	1.82	1.98	5.63	1.70	1.93	2.01	1.81	2.11	5.15	0.93	0.94
	RMSE	2.59	0.55	0.60	2.59	0.54	0.61	0.67	0.57	0.68	2.58	0.27	0.26
	SD	0.95	0.55	0.55	0.95	0.54	0.57	0.62	0.55	0.61	0.83	0.27	0.26
IonicHB (96)	MAE	4.50	1.62	1.82	4.50	1.53	1.67	2.72	1.57	1.71	2.94	0.52	0.38
	MSE	-4.34	-1.51	-1.72	-4.34	-1.38	-1.53	-2.38	-1.42	-1.57	-2.94	-0.51	-0.33
	MAXE	14.62	6.49	7.02	14.62	6.23	6.97	10.73	6.48	7.35	5.28	1.41	1.03
	RMSE	5.82	2.11	2.31	5.82	1.98	2.16	3.79	2.04	2.24	3.13	0.60	0.45
	SD	3.91	1.47	1.55	3.91	1.42	1.53	2.97	1.47	1.61	1.07	0.32	0.31
HB375x10 (3749)	MAE	1.68	0.49	0.46	1.68	0.51	0.50	0.53	0.51	0.51	1.81	0.29	0.28
	MSE	-1.62	0.25	0.13	-1.62	0.28	0.18	0.18	0.23	0.12	-1.81	0.07	0.06
	MAXE	7.49	8.13	7.51	7.49	8.23	7.56	5.85	8.08	7.32	6.11	2.62	2.69
	RMSE	2.29	0.76	0.73	2.29	0.79	0.77	0.77	0.78	0.78	2.11	0.42	0.40
	SD	1.62	0.72	0.72	1.62	0.74	0.75	0.75	0.75	0.77	1.08	0.41	0.40
IHB100x10 (350)	MAE	4.72	1.92	1.96	4.72	1.95	1.99	2.72	2.01	2.07	2.25	0.74	0.72
	MSE	-4.19	-0.40	-0.74	-4.19	-0.33	-0.63	-1.71	-0.41	-0.74	-1.98	-0.29	-0.24
	MAXE	17.17	7.46	7.48	17.17	8.72	8.56	11.84	9.02	9.11	4.79	3.34	3.14
	RMSE	6.18	2.48	2.56	6.18	2.53	2.63	3.77	2.62	2.77	2.59	0.99	0.97
	SD	4.56	2.45	2.45	4.56	2.52	2.56	3.37	2.59	2.67	1.66	0.95	0.94

Dataset (# of data points)		MINIs	MINIs-ACP (fit)	MINIs-ACP (scf)	MINIX	MINIX-ACP (fit)	MINIX-ACP (scf)	HF-3c	HF-3c-ACP (fit)	HF3c-ACP (scf)	6-31G*	6-31G*-ACP (fit)	6-31G*-ACP (scf)
HB300SPXx10 (1980)	MAE	1.73	1.04	1.03	1.44	0.78	0.76	1.98	0.92	0.89	1.30	0.45	0.44
	MSE	-0.80	0.22	0.01	-0.69	0.18	-0.01	1.69	0.28	0.07	-1.26	0.12	0.04
	MAXE	19.89	17.37	19.57	19.39	12.44	14.30	21.15	13.27	15.11	7.67	4.81	4.05
	RMSE	2.99	1.72	1.77	2.50	1.38	1.42	3.35	1.57	1.60	1.79	0.67	0.65
	SD	2.88	1.70	1.77	2.41	1.37	1.42	2.90	1.54	1.60	1.28	0.66	0.65
Pisub (105)	MAE	2.03	0.54	0.55	2.02	0.56	0.55	0.57	0.60	0.58	1.02	0.26	0.35
	MSE	-2.03	0.31	0.31	-2.02	0.34	0.32	-0.36	0.41	0.39	-1.02	-0.10	-0.27
	MAXE	4.77	2.58	2.62	4.77	2.76	2.77	1.62	2.75	2.75	3.26	0.72	0.90
	RMSE	2.19	0.75	0.75	2.17	0.79	0.78	0.69	0.83	0.82	1.20	0.31	0.40
	SD	0.83	0.68	0.69	0.81	0.71	0.71	0.60	0.72	0.72	0.64	0.30	0.30
Pi29n (29)	MAE	0.89	0.35	0.36	0.84	0.30	0.30	0.29	0.32	0.31	0.44	0.23	0.27
	MSE	-0.83	0.30	0.33	-0.84	0.27	0.27	-0.01	0.29	0.28	-0.36	-0.01	-0.12
	MAXE	1.98	2.51	2.50	2.86	1.66	1.56	1.44	2.08	2.01	4.19	0.91	1.47
	RMSE	1.01	0.59	0.60	1.02	0.43	0.42	0.38	0.49	0.48	0.91	0.32	0.40
	SD	0.59	0.51	0.51	0.59	0.34	0.32	0.39	0.40	0.39	0.86	0.33	0.39
BzDC215 (170)	MAE	0.92	0.37	0.40	0.85	0.37	0.39	0.23	0.42	0.44	1.27	0.29	0.34
	MSE	-0.90	-0.18	-0.21	-0.82	-0.18	-0.21	-0.03	-0.24	-0.26	-1.27	-0.24	-0.30
	MAXE	4.22	2.08	2.15	4.22	2.18	2.26	1.45	2.27	2.68	4.79	1.45	1.66
	RMSE	1.39	0.60	0.65	1.30	0.61	0.65	0.36	0.69	0.74	1.70	0.42	0.49
	SD	1.06	0.57	0.62	1.02	0.58	0.62	0.36	0.65	0.69	1.13	0.35	0.38
Hill18 (18)	MAE	2.84	0.65	0.67	1.90	0.53	0.61	2.30	0.58	0.61	1.92	0.58	0.59
	MSE	1.10	0.39	0.42	-0.31	0.22	0.19	2.30	0.31	0.25	-1.05	0.12	0.05
	MAXE	22.12	2.70	2.54	12.90	1.96	1.87	17.58	2.17	2.09	7.77	2.18	2.53
	RMSE	5.60	0.92	0.93	3.36	0.70	0.76	4.54	0.79	0.81	2.45	0.77	0.82
	SD	5.65	0.86	0.85	3.44	0.69	0.75	4.03	0.75	0.80	2.28	0.78	0.84
X40x10 (220)	MAE	1.53	0.61	0.63	1.41	0.56	0.58	0.86	0.62	0.63	1.14	0.36	0.36
	MSE	-1.29	0.00	-0.11	-1.35	-0.08	-0.18	0.37	-0.02	-0.14	-1.14	0.06	-0.01
	MAXE	9.15	5.02	5.04	8.13	5.06	5.15	7.43	4.87	5.08	4.31	1.81	2.13
	RMSE	2.67	1.03	1.05	2.42	1.00	1.02	1.39	1.03	1.06	1.60	0.49	0.52
	SD	2.35	1.03	1.05	2.01	1.00	1.01	1.34	1.04	1.06	1.13	0.49	0.52
PNICO23 (23)	MAE	1.96	0.73	0.68	2.17	0.71	0.67	1.80	0.70	0.72	1.64	0.36	0.34
	MSE	0.10	-0.13	0.09	1.74	0.08	0.36	0.28	0.09	0.43	1.28	0.02	0.17
	MAXE	17.97	2.35	2.41	9.02	2.76	3.09	10.08	2.58	2.94	5.34	1.54	1.55
	RMSE	4.07	0.96	0.91	3.22	0.98	1.03	3.01	0.95	1.03	2.29	0.50	0.49
	SD	4.16	0.97	0.92	2.77	1.00	0.99	3.07	0.97	0.96	1.95	0.52	0.47
CARBHB12 (12)	MAE	1.71	0.80	0.79	1.10	0.87	0.94	0.68	0.91	0.92	1.66	0.68	0.66
	MSE	1.67	0.21	0.17	1.10	-0.18	-0.32	-0.09	-0.09	-0.21	1.66	0.35	0.30
	MAXE	5.35	2.70	2.57	3.69	2.59	2.59	2.10	2.93	2.73	3.38	1.75	1.69
	RMSE	2.37	1.10	1.08	1.59	1.17	1.22	0.92	1.20	1.22	1.91	0.83	0.79
	SD	1.76	1.13	1.11	1.20	1.21	1.23	0.96	1.25	1.25	0.99	0.78	0.76
ADIM6 (6)	MAE	1.90	0.14	0.13	1.90	0.19	0.18	0.47	4.32	0.19	1.15	0.28	0.12
	MSE	-1.90	0.14	0.13	-1.90	0.19	0.18	-0.47	4.32	0.19	-1.15	0.28	0.12
	MAXE	3.16	0.25	0.23	3.16	0.32	0.31	0.71	6.60	0.27	1.86	0.59	0.32
	RMSE	2.09	0.16	0.15	2.09	0.21	0.20	0.52	4.57	0.20	1.24	0.33	0.16
	SD	0.95	0.07	0.07	0.95	0.08	0.07	0.24	1.62	0.06	0.52	0.18	0.11
HC12 (12)	MAE	1.80	0.20	0.20	1.80	0.23	0.22	0.42	0.26	0.25	1.19	0.36	0.27
	MSE	-1.80	0.10	0.09	-1.80	0.15	0.14	-0.42	0.19	0.18	-1.19	0.09	-0.07
	MAXE	3.39	0.49	0.51	3.39	0.45	0.46	0.91	0.53	0.53	2.54	0.60	0.78
	RMSE	1.96	0.25	0.26	1.96	0.27	0.26	0.51	0.29	0.29	1.30	0.38	0.34
	SD	0.81	0.24	0.25	0.81	0.23	0.23	0.30	0.23	0.24	0.55	0.38	0.34
HW30 (30)	MAE	0.81	0.27	0.27	0.81	0.29	0.29	0.31	0.28	0.29	1.49	0.32	0.34
	MSE	-0.80	0.02	-0.02	-0.80	-0.03	-0.06	0.05	-0.06	-0.09	-1.49	-0.07	-0.08

Dataset (# of data points)		MINIs	MINIs-ACP (fit)	MINIs-ACP (scf)	MINIX	MINIX-ACP (fit)	MINIX-ACP (scf)	HF-3c	HF-3c-ACP (fit)	HF3c-ACP (scf)	6-31G*	6-31G*-ACP (fit)	6-31G*-ACP (scf)
	MAXE	1.86	0.76	0.80	1.86	0.77	0.80	0.82	0.88	0.90	3.63	0.92	0.96
	RMSE	0.95	0.33	0.34	0.95	0.35	0.36	0.39	0.36	0.37	1.73	0.38	0.40
	SD	0.53	0.34	0.35	0.53	0.36	0.36	0.39	0.36	0.37	0.89	0.38	0.40
C2H4NT (75)	MAE	1.71	0.50	0.51	1.71	0.49	0.50	0.48	0.53	0.53	0.73	0.15	0.19
	MSE	-1.70	-0.08	-0.10	-1.70	-0.11	-0.12	-0.45	-0.10	-0.11	-0.73	-0.03	-0.18
	MAXE	5.52	2.02	2.09	5.52	2.11	2.16	2.06	2.14	2.19	1.62	0.52	0.71
	RMSE	2.53	0.74	0.75	2.53	0.73	0.74	0.70	0.77	0.78	0.86	0.19	0.23
	SD	1.89	0.74	0.75	1.89	0.73	0.74	0.54	0.77	0.78	0.46	0.18	0.14
CH4PAH (382)	MAE	1.19	0.18	0.17	1.19	0.17	0.17	0.19	0.16	0.16	0.74	0.14	0.11
	MSE	-1.19	0.15	0.14	-1.19	0.13	0.12	-0.19	0.06	0.05	-0.74	0.07	-0.05
	MAXE	6.94	0.78	0.77	6.94	0.74	0.71	1.58	0.68	0.73	4.17	1.01	1.68
	RMSE	1.84	0.25	0.24	1.84	0.24	0.23	0.32	0.20	0.21	1.00	0.20	0.22
	SD	1.40	0.19	0.19	1.40	0.20	0.20	0.26	0.20	0.20	0.68	0.19	0.21
CO2MOF (20)	MAE	2.19	0.79	0.86	2.18	0.75	0.81	0.92	0.79	0.85	2.55	0.77	0.78
	MSE	-2.15	-0.69	-0.78	-2.14	-0.67	-0.75	-0.71	-0.71	-0.79	-2.55	-0.64	-0.68
	MAXE	4.45	2.43	2.50	4.45	2.35	2.41	2.41	2.33	2.39	4.29	2.08	2.13
	RMSE	2.51	1.07	1.15	2.50	1.01	1.09	1.27	1.06	1.15	2.79	0.95	0.97
	SD	1.32	0.83	0.87	1.32	0.78	0.82	1.08	0.82	0.85	1.16	0.72	0.71
CO2PAH (249)	MAE	1.64	0.42	0.43	1.64	0.43	0.44	0.55	0.41	0.42	0.97	0.31	0.32
	MSE	-1.59	0.14	0.13	-1.59	0.15	0.13	-0.32	0.14	0.12	-0.90	0.09	0.01
	MAXE	9.01	2.09	2.10	9.01	2.15	2.12	2.70	1.93	1.89	5.44	1.56	1.65
	RMSE	2.62	0.56	0.57	2.62	0.59	0.59	0.82	0.55	0.55	1.47	0.43	0.44
	SD	2.09	0.54	0.55	2.09	0.58	0.58	0.76	0.54	0.54	1.16	0.42	0.44
CO2NPHAC (96)	MAE	1.92	0.39	0.45	1.92	0.33	0.38	0.63	0.35	0.41	1.38	0.30	0.27
	MSE	-1.92	-0.09	-0.16	-1.92	0.00	-0.08	-0.58	0.04	-0.04	-1.38	0.06	0.03
	MAXE	11.42	3.19	3.59	11.42	2.69	3.18	6.59	2.68	3.25	4.10	1.71	1.58
	RMSE	3.04	0.59	0.68	3.04	0.53	0.60	1.19	0.59	0.66	1.81	0.43	0.39
	SD	2.37	0.58	0.66	2.37	0.53	0.59	1.05	0.59	0.66	1.17	0.43	0.39
BzGas (129)	MAE	0.92	0.37	0.38	0.92	0.38	0.39	0.35	0.37	0.38	0.63	0.21	0.19
	MSE	-0.88	0.02	0.01	-0.88	0.01	0.00	-0.21	0.01	0.00	-0.57	0.06	0.00
	MAXE	6.13	2.63	2.58	6.13	2.94	2.84	3.01	3.19	3.09	1.54	1.20	1.14
	RMSE	1.41	0.55	0.56	1.41	0.58	0.58	0.60	0.58	0.58	0.73	0.28	0.26
	SD	1.11	0.55	0.56	1.11	0.58	0.58	0.56	0.59	0.59	0.45	0.27	0.27
Water38 (38)	MAE	30.62	0.70	1.56	30.62	0.62	1.24	7.67	0.70	1.32	19.51	0.64	0.77
	MSE	-30.62	-0.56	-1.49	-30.62	-0.45	-1.14	-7.67	-0.55	-1.22	-19.51	-0.54	0.61
	MAXE	60.84	1.70	2.79	60.84	1.70	2.23	15.61	1.65	2.25	38.12	1.84	1.69
	RMSE	33.29	0.85	1.70	33.29	0.77	1.36	8.53	0.85	1.44	21.17	0.78	0.88
	SD	13.25	0.65	0.84	13.25	0.63	0.75	3.77	0.65	0.77	8.33	0.57	0.65
Water1888 (1888)	MAE	1.44	0.65	0.66	1.44	0.64	0.65	0.70	0.65	0.66	1.48	0.42	0.41
	MSE	-1.32	-0.04	-0.06	-1.32	-0.02	-0.04	0.03	0.00	-0.02	-1.34	0.01	0.02
	MAXE	5.92	3.94	4.17	5.92	4.01	4.18	5.01	4.21	4.38	4.53	1.85	1.93
	RMSE	1.85	0.85	0.86	1.85	0.83	0.84	0.91	0.85	0.85	1.75	0.57	0.56
	SD	1.30	0.85	0.86	1.30	0.83	0.84	0.91	0.85	0.85	1.12	0.57	0.56
Water-2body (410)	MAE	0.63	0.14	0.14	0.63	0.14	0.14	0.17	0.15	0.14	0.65	0.08	0.09
	MSE	-0.56	0.06	0.04	-0.56	0.07	0.05	-0.02	0.07	0.06	-0.59	-0.02	0.00
	MAXE	3.59	0.70	0.70	3.59	0.70	0.66	0.85	0.76	0.75	3.27	0.91	0.87
	RMSE	1.26	0.21	0.21	1.26	0.21	0.20	0.25	0.21	0.21	1.12	0.14	0.15
	SD	1.13	0.20	0.20	1.13	0.20	0.20	0.25	0.20	0.20	0.96	0.14	0.15
B-set (160)	MAE	3.84	0.85	0.84	3.55	0.75	0.87	2.55	0.73	0.95	1.70	0.44	0.45
	MSE	-2.65	0.00	-0.25	-2.99	-0.09	-0.34	-1.07	-0.09	-0.52	-1.37	-0.12	-0.22
	MAXE	32.10	5.86	6.97	16.77	4.59	5.55	24.77	5.03	5.96	7.75	2.53	2.55
	RMSE	6.56	1.42	1.48	5.77	1.20	1.46	4.98	1.18	1.65	2.41	0.65	0.67

Dataset (# of data points)		MINIs	MINIs-ACP (fit)	MINIs-ACP (scf)	MINIX	MINIX-ACP (fit)	MINIX-ACP (scf)	HF-3c	HF-3c-ACP (fit)	HF3c-ACP (scf)	6-31G*	6-31G*-ACP (fit)	6-31G*-ACP (scf)
	SD	6.02	1.42	1.46	4.95	1.20	1.42	4.88	1.18	1.58	1.98	0.64	0.64
F-set (160)	MAE	1.37	0.49	0.44	1.37	0.50	0.45	1.37	0.64	0.56	1.08	0.29	0.29
	MSE	-1.11	0.25	0.14	-1.11	0.23	0.11	1.22	0.33	0.20	-0.91	0.05	-0.01
	MAXE	6.74	6.34	7.16	6.74	4.56	5.61	14.90	5.42	5.98	4.25	1.65	2.21
	RMSE	2.03	0.86	0.82	2.03	0.80	0.75	2.69	0.98	0.89	1.47	0.43	0.43
	SD	1.71	0.83	0.81	1.71	0.77	0.75	2.41	0.92	0.87	1.16	0.43	0.44
Si-set (152)	MAE	2.43	0.65	0.63	2.39	0.48	0.46	1.19	0.49	0.49	1.99	0.47	0.46
	MSE	-0.09	-0.30	-0.47	-2.37	-0.19	-0.37	-0.91	-0.17	-0.38	-1.99	-0.16	-0.22
	MAXE	20.52	3.28	3.77	13.71	3.97	3.60	8.68	3.98	3.36	9.74	3.84	3.91
	RMSE	4.52	1.00	0.99	4.13	0.87	0.77	2.30	0.89	0.81	3.06	0.75	0.75
	SD	4.53	0.95	0.88	3.39	0.85	0.68	2.12	0.88	0.72	2.33	0.73	0.72
P-set (120)	MAE	1.90	0.70	0.72	1.26	0.56	0.59	1.04	0.61	0.63	1.01	0.37	0.40
	MSE	-1.29	-0.30	-0.39	-0.48	0.22	0.14	0.64	0.24	0.15	-0.91	0.13	0.14
	MAXE	17.96	5.88	6.16	15.53	3.36	3.28	6.58	3.34	3.31	5.48	1.82	2.12
	RMSE	3.51	1.22	1.25	2.56	0.89	0.94	1.63	0.93	0.99	1.51	0.55	0.62
	SD	3.28	1.19	1.20	2.52	0.86	0.93	1.51	0.90	0.98	1.20	0.54	0.61
	MAE	0.68	0.49	0.49	0.50	0.35	0.36	0.70	0.41	0.08	0.79	0.25	0.26
S-set (144)	MSE	-0.47	0.12	0.14	-0.44	0.05	0.05	0.51	0.07	0.42	-0.77	0.06	0.03
	MAXE	4.25	3.35	3.30	4.36	2.95	2.98	6.57	3.19	3.20	2.55	1.45	1.74
	RMSE	1.03	0.77	0.76	0.87	0.53	0.54	1.30	0.62	0.63	1.03	0.35	0.38
	SD	0.92	0.76	0.75	0.76	0.53	0.54	1.20	0.61	0.63	0.69	0.34	0.38
Cl-set (160)	MAE	1.66	0.64	0.62	1.20	0.53	0.52	0.81	0.65	0.63	0.91	0.37	0.40
	MSE	-1.33	-0.07	-0.06	-1.07	-0.05	-0.09	0.26	0.03	-0.02	-0.37	0.26	0.22
	MAXE	11.83	6.14	7.03	9.01	4.19	5.26	7.82	5.16	5.77	5.40	2.72	2.99
	RMSE	3.02	1.02	1.04	2.31	0.81	0.82	1.33	0.96	0.94	1.44	0.62	0.70
	SD	2.72	1.02	1.04	2.06	0.82	0.81	1.31	0.97	0.95	1.40	0.56	0.66
SSI-anionic (575)	MAE	3.69	1.99	2.06	3.68	2.05	2.09	2.83	2.07	2.12	3.06	1.56	1.47
	MSE	-2.46	0.55	0.40	-2.45	0.47	0.37	-0.48	0.39	0.30	-2.75	-0.96	-0.95
	MAXE	18.25	8.18	8.14	18.25	8.20	8.25	12.06	8.05	8.41	9.18	7.07	6.83
	RMSE	5.19	2.47	2.56	5.18	2.54	2.60	3.56	2.56	2.62	3.82	2.09	2.00
	SD	4.57	2.41	2.53	4.57	2.50	2.57	3.53	2.53	2.61	2.66	1.85	1.76
WatAA-anionic (64)	MAE	6.33	1.47	1.65	6.33	1.57	1.69	3.18	1.69	1.81	4.08	1.01	0.86
	MSE	-6.33	-1.32	-1.55	-6.33	-1.46	-1.58	-3.09	-1.58	-1.71	-4.08	-1.00	-0.84
	MAXE	8.75	3.69	3.92	8.75	3.82	3.97	5.12	3.96	4.11	6.97	1.79	1.61
	RMSE	6.59	1.67	1.88	6.59	1.81	1.93	3.51	1.93	2.06	4.15	1.12	0.98
	SD	1.84	1.04	1.07	1.84	1.07	1.12	1.66	1.11	1.16	0.78	0.53	0.51
HSG-anionic (4)	MAE	4.88	1.78	1.82	4.88	1.89	1.92	2.47	2.02	2.06	4.42	0.93	0.85
	MSE	-4.56	-0.74	-0.82	-4.56	-0.82	-0.85	-1.27	-0.89	-0.91	-4.42	-0.93	-0.85
	MAXE	10.29	3.40	3.48	10.29	3.71	3.77	4.23	3.90	3.96	6.84	1.49	1.28
	RMSE	6.45	2.02	2.08	6.45	2.18	2.21	2.80	2.32	2.36	4.92	1.07	0.98
	SD	5.27	2.17	2.21	5.27	2.33	2.36	2.88	2.47	2.51	2.50	0.60	0.57
PLF547-anionic (155)	MAE	2.58	1.26	1.27	2.57	1.29	1.27	1.78	1.32	1.30	2.09	0.91	0.86
	MSE	-2.17	-0.24	-0.30	-2.14	-0.26	-0.29	-0.74	-0.28	-0.30	-1.95	-0.55	-0.56
	MAXE	24.85	10.82	11.19	23.84	10.34	10.49	16.83	10.41	10.47	11.58	7.21	6.99
	RMSE	5.46	2.37	2.43	5.41	2.43	2.44	3.71	2.49	2.49	3.45	1.74	1.68
	SD	5.03	2.37	2.42	4.99	2.42	2.43	3.64	2.48	2.48	2.86	1.65	1.59
IonicHB-anionic (24)	MAE	4.23	1.56	1.66	4.23	1.66	1.72	2.21	1.75	1.82	4.38	1.73	1.59
	MSE	-4.03	-1.17	-1.29	-4.03	-1.24	-1.31	-1.82	-1.34	-1.41	-4.38	-1.72	-1.57
	MAXE	11.80	4.62	4.93	11.80	4.87	5.06	6.25	4.97	5.17	8.09	2.84	2.56
	RMSE	5.53	2.08	2.21	5.53	2.22	2.30	2.94	2.34	2.42	4.89	1.91	1.75
	SD	3.87	1.76	1.83	3.87	1.89	1.93	2.36	1.95	2.01	2.24	0.85	0.79
	MAE	6.51	3.62	3.79	6.51	3.70	3.90	4.58	3.78	4.02	4.20	2.47	2.40

Dataset (# of data points)		MINIs	MINIs-ACP (fit)	MINIs-ACP (scf)	MINIX	MINIX-ACP (fit)	MINIX-ACP (scf)	HF-3c	HF-3c-ACP (fit)	HF3c-ACP (scf)	6-31G*	6-31G*-ACP (fit)	6-31G*-ACP (scf)
IHB100x10-anionic (650)	MSE	-6.19	-3.07	-3.30	-6.19	-3.12	-3.36	-4.04	-3.24	-3.51	-4.20	-2.44	-2.37
	MAXE	35.49	25.19	25.39	35.49	25.69	25.80	32.36	25.67	26.05	16.20	11.12	10.99
	RMSE	9.20	5.86	6.06	9.20	5.97	6.23	7.12	6.05	6.38	4.95	3.24	3.19
	SD	6.81	4.99	5.08	6.81	5.09	5.25	5.87	5.12	5.33	2.63	2.14	2.13
Ionic43-anionic (37)	MAE	10.77	4.27	5.13	10.85	4.39	5.08	6.70	4.35	4.97	4.26	2.55	2.56
	MSE	-10.77	-4.10	-4.99	-10.85	-4.12	-4.83	-6.12	-4.10	-4.78	-3.97	-2.21	-2.28
	MAXE	46.30	25.43	31.40	46.30	25.76	30.42	42.06	26.50	31.56	18.73	11.35	11.78
	RMSE	14.57	7.34	8.55	14.55	7.50	8.46	11.51	7.63	8.64	6.05	4.04	4.11
	SD	9.95	6.18	7.04	9.82	6.35	7.03	9.88	6.52	7.30	4.63	3.43	3.47
PEPCONF-Dipeptide (875)	MAE	1.85	1.01	1.01	1.84	0.96	0.95	1.26	1.01	1.00	1.28	0.51	0.54
	MSE	0.81	-0.07	-0.07	0.76	-0.04	-0.05	-0.05	-0.08	-0.10	1.09	0.27	0.26
	MAXE	11.39	5.00	5.48	11.39	4.45	4.73	7.60	4.44	4.70	5.32	2.03	2.05
	RMSE	2.44	1.29	1.29	2.41	1.24	1.22	1.78	1.28	1.28	1.62	0.66	0.68
	SD	2.30	1.29	1.29	2.29	1.24	1.22	1.78	1.28	1.28	1.20	0.60	0.63
TPCONF (8)	MAE	5.14	0.36	0.42	5.14	0.33	0.45	3.64	0.82	0.49	2.15	0.23	0.40
	MSE	-5.14	-0.21	-0.25	-5.14	-0.20	-0.31	-3.64	-0.71	-0.40	-1.63	-0.12	-0.40
	MAXE	11.97	1.70	1.84	11.97	1.65	1.89	8.23	1.65	2.00	3.74	0.38	0.67
	RMSE	6.49	0.63	0.69	6.49	0.61	0.73	4.45	1.00	0.79	2.37	0.25	0.45
	SD	4.24	0.64	0.69	4.24	0.62	0.70	2.73	0.76	0.73	1.84	0.24	0.22
P76 (71)	MAE	2.07	1.02	0.88	2.07	0.94	0.83	1.24	0.97	0.85	1.04	0.40	0.40
	MSE	2.02	0.06	0.18	2.02	0.08	0.11	0.63	0.07	0.11	0.20	0.09	0.03
	MAXE	7.36	3.67	3.75	7.36	3.29	3.09	4.60	3.21	3.03	3.74	1.86	1.94
	RMSE	2.65	1.28	1.17	2.65	1.19	1.09	1.67	1.22	1.09	1.38	0.55	0.55
	SD	1.72	1.29	1.16	1.72	1.20	1.09	1.55	1.22	1.09	1.37	0.55	0.56
YMPJ (495)	MAE	1.73	0.98	0.97	1.74	0.97	1.00	2.32	1.00	1.04	0.80	0.44	0.57
	MSE	-1.09	-0.36	-0.29	-1.17	-0.28	-0.35	-2.27	-0.34	-0.44	0.57	-0.33	-0.49
	MAXE	6.29	3.61	3.49	6.29	3.66	3.46	6.90	3.78	3.70	2.89	1.88	2.02
	RMSE	2.14	1.24	1.22	2.15	1.23	1.25	2.70	1.26	1.30	1.01	0.54	0.69
	SD	1.84	1.19	1.18	1.81	1.20	1.20	1.48	1.21	1.22	0.83	0.43	0.48
SPS (17)	MAE	1.49	0.63	0.64	1.69	0.61	0.57	0.58	0.57	0.53	0.84	0.25	0.21
	MSE	1.15	-0.13	-0.22	1.67	0.28	0.24	0.49	0.33	0.27	0.71	0.18	0.05
	MAXE	3.96	1.71	1.85	4.77	1.78	1.72	2.63	1.81	1.72	1.69	0.58	0.68
	RMSE	1.74	0.76	0.81	2.05	0.76	0.73	0.91	0.74	0.70	0.93	0.30	0.27
	SD	1.35	0.77	0.80	1.22	0.73	0.71	0.79	0.68	0.66	0.62	0.24	0.28
rSPS (45)	MAE	1.24	0.97	0.99	1.02	0.73	0.76	1.14	0.74	0.77	0.65	0.46	0.43
	MSE	-0.96	-0.78	-0.85	-0.57	-0.50	-0.57	-0.92	-0.53	-0.60	0.29	-0.29	-0.24
	MAXE	4.23	2.37	2.54	3.14	1.97	2.04	2.84	1.96	2.00	2.60	1.21	1.35
	RMSE	1.66	1.15	1.19	1.29	0.89	0.91	1.33	0.89	0.91	0.83	0.55	0.51
	SD	1.37	0.85	0.84	1.17	0.75	0.72	0.97	0.72	0.69	0.78	0.47	0.46
UpU46 (45)	MAE	2.17	1.39	1.30	3.18	1.68	1.61	2.94	1.66	1.57	1.67	0.88	0.75
	MSE	1.68	1.09	1.00	3.06	1.57	1.51	2.84	1.54	1.45	0.75	0.64	0.40
	MAXE	5.94	4.49	4.59	6.76	5.01	5.04	6.12	5.09	5.04	5.16	3.49	3.07
	RMSE	2.70	1.71	1.65	3.69	2.02	1.97	3.35	2.00	1.92	2.16	1.14	0.97
	SD	2.14	1.33	1.33	2.09	1.28	1.27	1.80	1.28	1.26	2.05	0.95	0.90
SCONF (17)	MAE	5.20	0.54	0.59	5.20	0.51	0.54	1.47	0.52	0.57	1.57	0.56	0.64
	MSE	1.51	-0.48	-0.56	1.51	-0.45	-0.51	-0.42	-0.42	-0.48	0.76	-0.48	-0.57
	MAXE	14.76	2.98	3.32	14.76	2.72	3.24	7.63	2.82	3.39	3.43	0.99	1.22
	RMSE	6.35	0.89	1.00	6.35	0.82	0.96	2.57	0.85	1.01	1.76	0.62	0.74
	SD	6.35	0.77	0.85	6.35	0.71	0.84	2.61	0.76	0.91	1.64	0.41	0.48
DSCONF (27)	MAE	5.30	1.31	1.44	5.30	1.16	1.26	2.47	1.21	1.31	3.31	0.52	0.48
	MSE	2.32	0.52	0.68	2.32	0.46	0.54	-0.08	0.56	0.64	3.15	0.21	0.14
	MAXE	13.69	2.81	3.27	13.69	2.87	3.00	5.24	2.96	3.17	6.62	1.33	1.16

Dataset (# of data points)		MINIs	MINIs-ACP (fit)	MINIs-ACP (scf)	MINIX	MINIX-ACP (fit)	MINIX-ACP (scf)	HF-3c	HF-3c-ACP (fit)	HF3c-ACP (scf)	6-31G*	6-31G*-ACP (fit)	6-31G*-ACP (scf)
	RMSE	6.27	1.59	1.73	6.27	1.44	1.55	2.88	1.51	1.62	3.65	0.65	0.59
	SD	5.94	1.53	1.62	5.94	1.39	1.48	2.93	1.42	1.51	1.89	0.63	0.58
SacchCONF (56)	MAE	4.71	1.31	1.33	4.71	1.38	1.40	3.08	1.42	1.43	1.17	0.43	0.47
	MSE	2.74	-0.30	-0.26	2.74	-0.46	-0.45	0.81	-0.52	-0.51	0.94	0.21	0.22
	MAXE	21.05	3.95	3.96	21.05	4.51	4.48	12.77	4.68	4.62	4.15	2.18	2.17
	RMSE	7.49	1.57	1.60	7.49	1.68	1.68	4.62	1.73	1.72	1.47	0.59	0.64
	SD	7.03	1.55	1.59	7.03	1.63	1.63	4.59	1.66	1.66	1.15	0.56	0.61
CCONF (426)	MAE	5.28	1.27	1.28	5.28	1.34	1.36	3.79	1.40	1.44	1.20	0.73	0.76
	MSE	-1.34	-0.52	-0.43	-1.34	-0.63	-0.64	-1.51	-0.71	-0.75	0.71	-0.53	-0.53
	MAXE	17.62	5.37	5.26	17.62	5.88	5.90	10.90	5.92	6.00	4.56	3.50	3.66
	RMSE	6.84	1.67	1.66	6.84	1.77	1.80	4.62	1.85	1.89	1.53	0.91	0.94
	SD	6.71	1.59	1.61	6.71	1.66	1.69	4.37	1.71	1.74	1.36	0.74	0.78
ACONF (15)	MAE	1.44	0.19	0.18	1.44	0.16	0.17	0.89	0.15	0.15	0.18	0.13	0.16
	MSE	-1.44	-0.17	-0.16	-1.44	-0.13	-0.15	-0.89	-0.08	-0.11	-0.18	-0.11	-0.15
	MAXE	2.63	0.55	0.54	2.63	0.46	0.47	1.76	0.43	0.44	0.60	0.30	0.39
	RMSE	1.55	0.23	0.23	1.55	0.20	0.21	0.96	0.19	0.19	0.24	0.15	0.18
	SD	0.58	0.17	0.16	0.58	0.16	0.15	0.36	0.18	0.16	0.17	0.11	0.10
BCONF (64)	MAE	2.40	0.29	0.29	2.40	0.28	0.27	0.58	0.28	0.27	1.25	0.24	0.34
	MSE	2.34	-0.11	-0.11	2.34	-0.09	-0.07	0.53	-0.03	0.00	1.17	-0.17	-0.29
	MAXE	3.69	0.75	0.73	3.69	0.72	0.74	1.40	0.83	0.83	1.81	0.62	0.74
	RMSE	2.53	0.35	0.35	2.53	0.34	0.33	0.69	0.34	0.33	1.30	0.29	0.40
	SD	0.98	0.34	0.34	0.98	0.33	0.33	0.44	0.34	0.33	0.57	0.23	0.28
PentCONF (342)	MAE	0.96	0.16	0.16	0.96	0.19	0.15	0.55	0.27	0.21	0.47	0.19	0.15
	MSE	-0.96	0.10	0.10	-0.96	0.16	0.09	-0.54	0.26	0.19	0.46	0.17	0.12
	MAXE	1.94	1.08	1.03	1.94	1.21	1.00	1.27	1.62	1.35	1.77	0.72	0.50
	RMSE	1.07	0.29	0.28	1.07	0.33	0.27	0.65	0.45	0.38	0.58	0.24	0.19
	SD	0.49	0.28	0.26	0.49	0.29	0.26	0.36	0.37	0.33	0.36	0.16	0.14
Undecamer125 (124)	MAE	1.57	0.44	0.43	1.57	0.43	0.42	1.27	0.48	0.47	2.02	0.24	0.30
	MSE	1.57	-0.17	-0.08	1.57	-0.12	-0.05	0.32	-0.24	-0.15	0.93	-0.04	-0.06
	MAXE	3.43	1.48	1.65	3.43	1.57	1.69	3.87	1.53	1.67	4.89	0.96	1.22
	RMSE	1.69	0.55	0.54	1.69	0.53	0.52	1.54	0.61	0.59	2.31	0.32	0.39
	SD	0.63	0.53	0.53	0.63	0.52	0.52	1.51	0.56	0.57	2.12	0.32	0.39
ICONF (17)	MAE	2.56	0.93	0.94	1.74	0.96	0.93	2.29	1.01	0.97	1.00	0.34	0.36
	MSE	-2.06	-0.21	0.04	-1.12	-0.35	-0.30	-1.39	-0.36	-0.32	0.30	0.12	0.11
	MAXE	9.48	3.48	3.14	5.65	3.48	3.37	8.28	3.69	3.53	3.11	0.84	0.85
	RMSE	4.00	1.37	1.37	2.55	1.34	1.34	3.43	1.40	1.39	1.37	0.41	0.43
	SD	3.53	1.39	1.41	2.36	1.33	1.34	3.22	1.40	1.39	1.38	0.40	0.42
MCONF (51)	MAE	0.88	0.80	0.76	0.88	0.95	0.93	0.89	1.08	1.06	1.05	0.33	0.32
	MSE	0.44	-0.69	-0.63	0.44	-0.89	-0.86	-0.34	-1.03	-1.01	1.04	-0.10	-0.09
	MAXE	2.97	1.82	1.69	2.97	2.09	1.97	2.15	2.27	2.12	1.94	0.91	0.98
	RMSE	1.19	0.93	0.88	1.19	1.10	1.07	1.10	1.23	1.20	1.17	0.41	0.39
	SD	1.12	0.63	0.62	1.12	0.65	0.64	1.06	0.68	0.67	0.52	0.40	0.38
Torsion21 (189)	MAE	1.34	0.58	0.50	1.30	0.49	0.49	1.17	0.54	0.55	0.80	0.19	0.20
	MSE	0.47	-0.47	-0.32	0.42	-0.33	-0.30	0.53	-0.35	-0.33	0.72	0.01	0.03
	MAXE	4.78	2.19	2.06	4.39	1.76	1.77	3.53	1.95	1.97	2.47	0.61	0.62
	RMSE	1.73	0.78	0.68	1.65	0.64	0.63	1.50	0.69	0.70	0.97	0.23	0.25
	SD	1.67	0.63	0.61	1.59	0.55	0.56	1.40	0.60	0.61	0.64	0.23	0.25
37Conf8 (258)	MAE	2.66	1.61	1.63	2.48	1.46	1.50	1.90	1.47	1.50	1.13	0.75	0.77
	MSE	-0.58	-0.61	-0.64	-0.77	-0.75	-0.83	-1.29	-0.74	-0.82	0.49	-0.20	-0.28
	MAXE	16.36	9.30	9.66	16.36	9.38	9.68	10.70	9.18	9.47	7.77	3.68	4.13
	RMSE	3.78	2.31	2.36	3.49	2.07	2.14	2.67	2.06	2.13	1.61	1.08	1.15
	SD	3.74	2.23	2.28	3.42	1.93	1.98	2.34	1.93	1.97	1.53	1.07	1.11

Dataset (# of data points)		**MINIs**	**MINIs-ACP (fit)**	**MINIs-ACP (scf)**	**MINIX**	**MINIX-ACP (fit)**	**MINIX-ACP (scf)**	**HF-3c**	**HF-3c-ACP (fit)**	**HF3c-ACP (scf)**	**6-31G***	**6-31G*-ACP (fit)**	**6-31G*-ACP (scf)**
DCONF (2142)	MAE	1.11	0.64	0.64	1.05	0.60	0.60	0.88	0.61	0.62	0.59	0.28	0.29
	MSE	0.18	-0.13	-0.11	0.34	-0.01	-0.05	0.22	-0.02	-0.06	0.52	0.13	0.12
	MAXE	7.78	4.43	4.32	3.53	3.18	3.28	4.17	3.29	3.42	2.46	1.26	1.29
	RMSE	1.54	0.93	0.93	1.34	0.82	0.85	1.16	0.85	0.87	0.81	0.38	0.39
	SD	1.53	0.92	0.92	1.30	0.82	0.85	1.13	0.85	0.87	0.62	0.35	0.37
MolCONF (5623)	MAE	0.81	0.53	0.54	0.76	0.49	0.49	0.56	0.50	0.50	0.42	0.32	0.32
	MSE	-0.05	-0.31	-0.30	-0.03	-0.24	-0.24	-0.21	-0.25	-0.25	0.08	-0.07	-0.06
	MAXE	15.13	16.79	16.94	9.87	16.88	16.73	10.13	16.08	15.92	5.79	5.71	5.56
	RMSE	1.42	1.01	1.02	1.28	0.93	0.93	0.98	0.92	0.93	0.70	0.52	0.52
	SD	1.42	0.97	0.98	1.28	0.90	0.90	0.96	0.89	0.89	0.70	0.52	0.51
MOLdef (9298)	MAE	4.68	2.31	2.35	3.82	2.26	2.23	3.32	2.35	2.33	2.05	1.30	1.35
	MSE	2.65	0.73	0.74	2.83	0.87	0.79	2.89	0.96	0.88	1.75	0.97	1.02
	MAXE	99.27	50.54	58.34	79.65	48.44	53.66	73.23	48.17	52.92	29.73	32.90	33.17
	RMSE	8.57	3.99	4.17	5.94	3.78	3.78	4.79	3.90	3.90	2.95	2.09	2.12
	SD	8.15	3.92	4.11	5.22	3.68	3.69	3.82	3.78	3.80	2.38	1.85	1.86
MOLdef-H2O (990)	MAE	1.78	0.42	0.39	1.78	0.40	0.37	1.02	0.40	0.36	0.62	0.33	0.37
	MSE	0.04	-0.07	-0.08	0.04	-0.02	-0.03	0.24	0.00	0.00	0.39	-0.16	-0.17
	MAXE	10.06	4.93	4.12	10.06	4.74	4.04	5.82	4.87	4.11	2.95	3.04	3.45
	RMSE	2.58	0.69	0.62	2.58	0.67	0.61	1.40	0.68	0.60	0.92	0.48	0.55
	SD	2.58	0.69	0.62	2.58	0.67	0.61	1.38	0.68	0.60	0.83	0.46	0.52
ANI1ccxCONF (32944)	MAE	5.86	2.59	2.57	5.86	2.68	2.64	4.44	2.82	2.78	3.96	1.63	1.74
	MSE	2.70	1.12	1.03	2.70	1.18	1.00	3.74	1.37	1.20	3.51	1.00	1.02
	MAXE	69.05	31.29	31.86	69.05	56.32	32.48	38.86	64.11	33.58	38.12	16.66	17.10
	RMSE	8.35	3.77	3.77	8.35	3.91	3.82	6.34	4.12	4.01	5.47	2.33	2.45
	SD	7.90	3.60	3.63	7.90	3.72	3.69	5.12	3.88	3.83	4.20	2.10	2.23
PEPCONF-Dipeptide-anionic (175)	MAE	2.39	1.00	0.99	2.37	1.01	1.02	1.98	1.01	1.03	1.26	0.75	0.79
	MSE	-0.30	-0.25	-0.25	-0.34	-0.23	-0.25	-0.49	-0.24	-0.28	0.50	0.00	-0.02
	MAXE	7.97	3.46	3.48	7.97	3.45	3.48	5.69	3.45	3.52	4.08	2.41	2.61
	RMSE	2.92	1.23	1.23	2.89	1.25	1.27	2.37	1.25	1.29	1.60	0.92	0.96
	SD	2.92	1.21	1.21	2.88	1.23	1.24	2.32	1.23	1.26	1.52	0.92	0.97
MolCONF-anionic (79)	MAE	4.39	1.34	1.24	2.76	1.67	1.64	2.12	1.84	1.83	0.68	0.29	0.31
	MSE	4.10	0.79	0.63	2.49	0.26	0.25	0.61	0.28	0.25	-0.58	-0.12	-0.11
	MAXE	20.49	8.24	8.84	16.43	9.02	9.15	13.57	9.72	9.84	2.46	1.31	1.28
	RMSE	8.40	2.70	2.71	5.49	3.19	3.20	4.41	3.50	3.53	1.11	0.48	0.50
	SD	7.37	2.59	2.65	4.93	3.20	3.21	4.40	3.51	3.54	0.95	0.47	0.49

Table S4. Detailed error analysis with respect to reference data in the validation set. The numbers in bracket in the first column indicates the number of data points. The various shorthand notations are as follows: MINIs = HF-D3/MINIs, MINIs-ACP = HF-D3/MINIs-ACP, MINIX = HF-D3/MINIX, MINIX-ACP = HF-D3/MINIX-ACP, 6-31G* = HF-D3/6-31G*, 6-31G*-ACP = HF-D3/6-31G*-ACP, MAE = mean absolute error in kcal/mol, MSE = mean signed error in kcal/mol, MAXE = maximum absolute error in kcal/mol, RMSE = root-mean-square error in kcal/mol, and SD = standard deviation in kcal/mol.

Dataset (# of datapoints)		**MINIs**	**MINIs-ACP**	**MINIX**	**MINIX-ACP**	**HF-3c**	**HF-3c-ACP**	**6-31G***	**6-31G*-ACP**
BlindNCI (80)	MAE	1.00	0.34	1.00	0.36	0.38	0.35	1.10	0.25
	MSE	-0.97	-0.04	-0.97	0.01	0.09	0.02	-1.10	0.07
	MAXE	7.82	3.11	7.82	3.39	3.44	3.51	6.98	3.78
	RMSE	2.03	0.64	2.03	0.70	0.83	0.69	1.89	0.58
	SD	1.80	0.65	1.80	0.70	0.83	0.70	1.55	0.58
DES15K (11474)	MAE	2.78	1.36	2.66	1.25	2.14	1.32	1.80	0.57
	MSE	-2.26	0.10	-2.24	0.30	1.37	0.36	-1.65	0.11
	MAXE	26.07	25.65	21.86	16.62	26.11	20.97	8.57	8.83
	RMSE	4.08	2.44	3.91	2.22	4.02	2.37	2.17	0.92
	SD	3.40	2.44	3.20	2.21	3.78	2.35	1.41	0.92
NENCI-2021 (5859)	MAE	2.34	1.04	2.26	1.05	1.61	1.07	1.86	0.51
	MSE	-1.89	0.35	-1.84	0.52	0.93	0.57	-1.51	0.15
	MAXE	26.44	38.08	21.45	39.17	39.84	39.37	12.55	17.22
	RMSE	3.42	2.57	3.26	2.55	3.71	2.60	2.29	1.10
	SD	2.85	2.55	2.70	2.50	3.59	2.53	1.72	1.09
R160x6 (960)	MAE	0.81	0.79	0.81	0.81	0.67	0.79	0.70	0.79
	MSE	-0.57	0.37	-0.57	0.31	0.30	0.32	-0.01	0.42
	MAXE	6.20	6.96	6.20	7.13	9.57	7.39	6.26	6.60
	RMSE	1.12	1.17	1.12	1.18	1.10	1.17	1.00	1.15
	SD	0.96	1.11	0.96	1.14	1.06	1.13	1.00	1.07
R739x5 (4330)	MAE	0.91	0.85	0.72	0.63	0.90	0.66	0.56	0.46
	MSE	-0.08	0.14	-0.27	0.16	0.72	0.08	-0.13	0.13
	MAXE	13.47	7.60	13.65	7.67	21.27	9.53	4.18	3.39
	RMSE	1.31	1.19	1.20	0.88	1.82	0.94	0.77	0.62
	SD	1.31	1.18	1.17	0.86	1.68	0.94	0.76	0.60
CE20 (20)	MAE	16.64	1.63	16.64	1.55	3.32	1.84	11.10	1.31
	MSE	16.64	0.34	16.64	0.18	2.33	0.42	11.10	0.77
	MAXE	43.29	6.30	43.29	6.28	9.63	6.11	22.33	3.74
	RMSE	20.39	2.20	20.39	2.18	4.16	2.43	12.51	1.66
	SD	12.09	2.23	12.09	2.23	3.53	2.45	5.93	1.51
CHAL336 (48)	MAE	2.66	1.73	2.17	2.21	1.30	2.32	2.23	2.07
	MSE	1.46	-1.60	-1.78	-2.18	0.46	-2.29	-2.23	-2.04
	MAXE	24.06	12.19	11.47	20.10	11.65	19.37	14.07	12.63
	RMSE	5.88	2.69	2.80	3.71	2.67	3.70	2.91	3.25
	SD	5.76	2.19	2.19	3.04	2.66	2.93	1.89	2.56
XB45 (33)	MAE	12.53	2.99	6.66	2.64	7.33	2.66	2.01	2.78

Dataset (# of datapoints)		MINIs	MINIs-ACP	MINIX	MINIX-ACP	HF-3c	HF-3c-ACP	6-31G*	6-31G*-ACP
	MSE	-11.64	0.38	-3.19	0.34	-6.04	0.61	1.53	2.66
	MAXE	77.33	16.31	32.32	10.63	38.25	10.12	6.39	5.62
	RMSE	25.63	4.37	11.51	3.63	13.62	3.64	2.67	3.16
	SD	23.19	4.42	11.23	3.67	12.40	3.64	2.22	1.73
WaterOrg (2376)	MAE	1.81	0.77	1.81	0.75	0.71	0.71	2.44	0.40
	MSE	-1.80	0.76	-1.80	0.73	0.69	0.68	-2.44	0.23
	MAXE	4.57	2.36	4.57	2.34	2.86	2.40	5.97	1.41
	RMSE	2.03	0.91	2.03	0.91	0.93	0.87	2.63	0.47
	SD	0.94	0.50	0.94	0.54	0.62	0.54	0.97	0.41
Water27 (27)	MAE	49.77	13.79	49.77	13.89	21.88	13.99	29.93	7.02
	MSE	49.77	13.43	49.77	13.36	21.88	13.41	29.93	4.76
	MAXE	119.82	39.73	119.82	40.21	46.79	39.99	92.86	23.47
	RMSE	60.29	19.78	60.29	20.01	26.34	20.02	39.00	10.61
	SD	34.68	14.81	34.68	15.17	14.95	15.15	25.47	9.67
HW6Cl (6)	MAE	11.51	3.01	12.20	2.72	4.03	1.01	14.57	0.93
	MSE	-11.51	-3.01	-12.20	-2.72	4.03	-1.01	-14.57	-0.70
	MAXE	22.00	4.57	22.83	3.98	5.32	1.91	26.31	1.61
	RMSE	13.42	3.21	14.09	2.89	4.27	1.11	16.94	1.04
	SD	7.56	1.20	7.74	1.07	1.53	0.51	9.47	0.85
HW6F (6)	MAE	61.61	32.08	61.61	30.83	44.13	31.57	38.08	22.92
	MSE	-61.61	-32.08	-61.61	-30.83	-44.13	-31.57	-38.08	-22.92
	MAXE	81.32	38.33	81.32	36.76	50.52	37.39	57.45	28.52
	RMSE	63.23	32.60	63.23	31.31	44.46	32.01	40.94	23.67
	SD	15.62	6.32	15.62	6.03	5.90	5.80	16.47	6.50
FmH2O10 (10)	MAE	107.13	39.49	107.13	37.48	50.73	37.45	85.57	31.36
	MSE	-107.13	-39.49	-107.13	-37.48	-50.73	-37.45	-85.57	-31.36
	MAXE	109.19	40.84	109.19	38.85	53.26	38.95	86.77	32.63
	RMSE	107.14	39.51	107.14	37.50	50.75	37.48	85.58	31.37
	SD	0.93	1.34	0.93	1.32	1.65	1.37	1.30	0.96
SW49Bind345 (30)	MAE	10.93	4.30	11.36	3.82	4.66	3.71	7.57	1.56
	MSE	-10.69	-3.49	-11.30	-3.28	-4.59	-3.09	-7.48	-1.56
	MAXE	19.77	11.24	20.56	9.81	8.79	9.82	13.48	2.90
	RMSE	13.22	5.13	13.69	4.62	5.61	4.47	9.08	1.76
	SD	7.92	3.83	7.86	3.30	3.28	3.29	5.22	0.83
SW49Bind6 (18)	MAE	27.56	9.82	27.96	8.71	11.20	8.40	18.85	3.24
	MSE	-27.56	-9.82	-27.96	-8.71	-11.20	-8.40	-18.85	-3.24
	MAXE	28.82	16.15	29.23	14.25	12.62	14.24	19.79	3.92
	RMSE	27.57	10.07	27.99	8.93	11.22	8.64	18.88	3.27
	SD	0.74	2.29	1.15	2.04	0.75	2.11	0.96	0.46
H2O20Bind10 (10)	MAE	110.97	6.53	110.97	4.80	16.90	4.14	103.54	2.00
	MSE	-110.97	-6.53	-110.97	-4.80	-16.90	-4.14	-103.54	-2.00
	MAXE	112.79	6.96	112.79	5.10	17.73	4.60	110.41	3.39
	RMSE	110.98	6.54	110.98	4.80	16.93	4.14	103.57	2.08

Dataset (# of datapoints)		MINIs	MINIs-ACP	MINIX	MINIX-ACP	HF-3c	HF-3c-ACP	6-31G*	6-31G*-ACP
	SD	1.12	0.27	1.12	0.23	1.06	0.22	2.72	0.61
L7 (7)	MAE	3.64	1.42	3.64	1.39	1.37	1.48	3.61	0.82
	MSE	-3.64	0.68	-3.64	0.81	0.50	0.94	-3.02	-0.32
	MAXE	6.82	3.73	6.82	3.82	2.69	4.44	10.39	2.04
	RMSE	4.38	1.78	4.38	1.79	1.54	1.97	4.73	1.07
	SD	2.63	1.78	2.63	1.73	1.57	1.87	3.93	1.10
S12L (10)	MAE	14.63	6.41	14.65	5.96	6.05	5.58	13.51	6.22
	MSE	-14.63	-6.13	-14.65	-5.60	-4.68	-4.94	-13.51	-6.09
	MAXE	25.18	16.92	25.18	16.09	11.75	15.88	21.01	15.05
	RMSE	15.84	8.72	15.84	8.25	7.55	7.86	14.58	7.66
	SD	6.40	6.53	6.36	6.39	6.24	6.45	5.77	4.90
S30L (26)	MAE	13.07	5.65	13.25	5.23	4.80	4.74	11.71	5.34
	MSE	-13.07	-4.55	-13.25	-3.91	-3.26	-3.24	-11.71	-4.05
	MAXE	41.59	24.32	41.59	21.80	10.23	22.02	38.15	12.48
	RMSE	14.92	7.90	15.05	7.33	5.75	7.04	13.52	6.74
	SD	7.33	6.59	7.29	6.33	4.84	6.37	6.90	5.50
C60dimer (14)	MAE	1.91	0.69	1.91	0.69	0.90	0.79	1.13	0.95
	MSE	-1.91	-0.33	-1.91	-0.27	-0.90	-0.06	-0.12	-0.79
	MAXE	7.61	1.64	7.61	1.70	3.04	2.27	2.86	1.85
	RMSE	2.64	0.81	2.64	0.80	1.10	0.92	1.36	1.11
	SD	1.90	0.77	1.90	0.78	0.67	0.95	1.41	0.81
Ni2021 (11)	MAE	25.53	5.60	25.53	5.89	6.74	5.98	22.01	10.05
	MSE	-23.69	3.82	-23.69	4.74	5.22	4.99	-20.37	9.73
	MAXE	99.52	14.11	99.45	18.82	21.05	20.43	57.33	49.23
	RMSE	35.82	7.05	35.81	8.17	8.93	8.52	26.05	16.45
	SD	28.18	6.21	28.16	6.98	7.59	7.24	17.04	13.92
Anionpi-anionic (16)	MAE	8.17	4.89	8.17	4.74	6.96	4.69	5.56	4.10
	MSE	-7.26	-2.29	-7.30	-1.91	-3.59	-2.24	-5.09	-2.78
	MAXE	48.58	29.59	48.58	27.75	40.33	27.00	19.08	13.95
	RMSE	14.67	8.82	14.67	8.33	11.81	8.27	7.91	5.67
	SD	13.17	8.80	13.15	8.37	11.62	8.22	6.25	5.10
IL236-anionic (236)	MAE	7.29	2.14	6.78	1.95	3.13	1.73	4.71	1.34
	MSE	-7.27	-0.58	-6.76	0.05	-0.60	-0.15	-4.63	-1.13
	MAXE	21.12	10.20	17.24	6.79	11.55	6.69	8.79	5.44
	RMSE	8.73	2.94	7.90	2.41	4.01	2.25	5.36	1.72
	SD	4.86	2.89	4.09	2.41	3.97	2.25	2.70	1.31
DES15K-anionic (1281)	MAE	7.87	4.04	7.40	4.14	6.89	4.50	4.07	2.27
	MSE	-6.11	-1.96	-5.70	-1.08	-0.49	-0.79	-3.03	-1.85
	MAXE	68.00	63.11	68.00	54.48	67.57	57.27	24.52	25.23
	RMSE	12.71	7.59	12.15	7.45	11.23	8.02	5.66	4.32
	SD	11.15	7.33	10.74	7.37	11.23	7.98	4.79	3.91
NENCI-2021-anionic (889)	MAE	8.66	6.15	8.25	6.21	7.20	6.55	5.17	3.27
	MSE	-4.59	-0.69	-4.36	-0.33	-1.37	-1.58	-2.35	-1.67

Dataset (# of datapoints)		**MINIs**	**MINIs-ACP**	**MINIX**	**MINIX-ACP**	**HF-3c**	**HF-3c-ACP**	**6-31G***	**6-31G*-ACP**
	MAXE	62.79	62.20	62.79	63.60	64.39	74.02	22.42	18.38
	RMSE	12.62	9.16	12.15	9.48	10.97	10.62	6.40	4.45
	SD	11.76	9.14	11.35	9.48	10.89	10.51	5.96	4.13
CHAL336-anionic (19)	MAE	8.51	12.32	15.66	16.24	14.10	16.98	5.86	5.92
	MSE	-1.33	-11.08	-15.23	-15.46	-12.57	-16.26	-4.48	-5.41
	MAXE	29.46	49.51	61.20	56.78	57.85	59.77	18.86	21.24
	RMSE	10.18	18.07	23.90	23.97	21.90	24.92	8.24	8.83
	SD	10.37	14.67	18.93	18.82	18.42	19.40	7.10	7.17
XB45-anionic (12)	MAE	25.07	31.12	33.13	30.68	31.40	31.84	15.64	15.48
	MSE	22.50	31.12	33.13	30.68	30.98	31.84	15.57	15.12
	MAXE	73.50	73.99	88.80	80.29	86.47	82.09	40.73	33.73
	RMSE	36.00	40.52	46.02	40.50	45.07	41.92	22.09	19.86
	SD	29.35	27.12	33.36	27.62	34.19	28.47	16.37	13.45
S30L-anionic (2)	MAE	31.48	8.27	31.48	6.53	15.50	7.19	14.60	5.45
	MSE	-31.48	-8.27	-31.48	-6.53	-15.50	-7.19	-14.60	-5.45
	MAXE	31.57	9.08	31.57	7.31	15.83	7.88	14.67	5.63
	RMSE	31.48	8.31	31.48	6.57	15.51	7.22	14.60	5.45
	SD	0.12	1.14	0.12	1.10	0.46	0.98	0.09	0.26
SafroleCONF (5)	MAE	0.92	0.75	0.92	0.72	0.82	0.70	0.44	0.45
	MSE	-0.92	-0.75	-0.92	-0.72	-0.82	-0.70	-0.44	-0.45
	MAXE	1.28	1.19	1.28	1.18	1.19	1.18	0.97	1.02
	RMSE	1.03	0.86	1.03	0.83	0.92	0.82	0.62	0.64
	SD	0.52	0.47	0.52	0.47	0.47	0.47	0.48	0.50
AlcoholCONF (31)	MAE	0.71	0.44	0.71	0.40	0.61	0.38	0.35	0.28
	MSE	-0.48	-0.17	-0.48	-0.07	-0.37	-0.03	0.31	0.11
	MAXE	2.30	1.17	2.30	1.04	2.05	0.96	0.90	0.67
	RMSE	0.89	0.51	0.89	0.48	0.77	0.47	0.41	0.34
	SD	0.76	0.49	0.76	0.49	0.68	0.47	0.28	0.33
BeranCONF (50)	MAE	1.59	0.61	1.54	0.67	1.03	0.71	0.65	0.35
	MSE	0.10	0.20	0.33	0.11	0.47	0.12	0.47	0.13
	MAXE	5.33	3.46	5.33	4.07	3.88	3.97	2.64	1.12
	RMSE	2.17	0.87	2.09	0.99	1.35	1.02	0.85	0.44
	SD	2.19	0.85	2.09	0.99	1.28	1.03	0.72	0.42
Torsion30 (2107)	MAE	1.62	0.54	1.62	0.58	1.18	0.59	0.59	0.44
	MSE	0.66	0.29	0.65	0.32	0.66	0.33	0.44	0.23
	MAXE	7.98	10.59	7.98	11.05	8.58	11.39	11.50	10.47
	RMSE	2.04	0.91	2.03	0.97	1.54	0.99	1.03	0.81
	SD	1.93	0.87	1.92	0.91	1.40	0.94	0.93	0.77
MPCONF196 (112)	MAE	3.90	1.82	3.90	1.74	2.86	1.84	3.51	1.01
	MSE	-0.23	-0.10	-0.23	-0.50	-1.82	-0.61	3.11	0.20
	MAXE	16.14	6.57	16.14	6.57	11.19	6.68	12.23	4.44
	RMSE	5.01	2.25	5.01	2.22	3.63	2.35	4.34	1.32
	SD	5.03	2.26	5.03	2.18	3.15	2.27	3.05	1.31

Dataset (# of datapoints)		MINIs	MINIs-ACP	MINIX	MINIX-ACP	HF-3c	HF-3c-ACP	6-31G*	6-31G*-ACP
PEPCONF-Tripeptide (647)	MAE	2.34	1.24	2.26	1.17	1.33	1.19	1.94	0.73
	MSE	1.32	-0.43	1.19	-0.42	-0.54	-0.44	1.58	0.06
	MAXE	10.01	5.04	10.01	4.67	6.25	4.52	7.18	3.10
	RMSE	2.92	1.57	2.83	1.48	1.70	1.51	2.42	0.92
	SD	2.61	1.51	2.58	1.42	1.62	1.44	1.84	0.92
PEPCONF-Disulfide (620)	MAE	3.77	2.71	3.71	2.58	2.64	2.61	2.53	2.21
	MSE	1.91	-1.35	1.84	-1.36	-0.83	-1.33	1.20	-1.36
	MAXE	15.43	15.90	16.00	15.86	16.40	16.04	9.56	15.19
	RMSE	4.79	3.60	4.71	3.48	3.56	3.50	3.15	3.30
	SD	4.39	3.34	4.34	3.20	3.46	3.23	2.92	3.01
PEPCONF-Cyclic (320)	MAE	3.18	1.95	3.18	2.01	3.23	2.04	4.62	1.44
	MSE	-1.13	1.20	-1.14	1.15	-1.65	0.93	4.54	1.00
	MAXE	15.38	7.45	15.38	7.37	12.49	7.60	14.47	6.96
	RMSE	4.17	2.50	4.18	2.59	4.07	2.62	5.35	1.81
	SD	4.03	2.20	4.02	2.33	3.73	2.45	2.83	1.51
PEPCONF-Bioactive (175)	MAE	3.58	1.80	3.58	1.75	2.46	1.77	1.82	0.86
	MSE	2.03	0.16	2.02	0.15	0.30	0.17	1.41	0.09
	MAXE	14.50	6.24	14.50	6.36	10.22	6.50	6.99	3.00
	RMSE	4.59	2.28	4.58	2.25	3.19	2.31	2.40	1.06
	SD	4.13	2.28	4.12	2.26	3.19	2.31	1.95	1.06
PEPCONF-Disulfide-anionic (150)	MAE	4.32	4.15	3.96	4.30	5.02	4.29	2.16	4.01
	MSE	-1.88	-3.42	-1.91	-3.76	-4.33	-3.73	-1.20	-3.59
	MAXE	19.50	15.78	18.97	16.01	21.68	16.04	7.33	16.38
	RMSE	5.58	5.57	5.30	5.70	6.80	5.71	2.73	5.23
	SD	5.27	4.41	4.96	4.29	5.25	4.33	2.47	3.82
PEPCONF-Bioactive-anionic (20)	MAE	2.99	2.20	2.99	2.16	2.70	2.09	1.40	0.84
	MSE	0.56	-0.76	0.56	-0.79	-0.59	-0.72	0.93	-0.21
	MAXE	11.00	5.42	11.00	5.52	9.60	5.62	4.14	2.52
	RMSE	4.16	2.67	4.16	2.62	3.51	2.56	1.76	1.06
	SD	4.23	2.63	4.23	2.56	3.55	2.52	1.53	1.06

Table S5. Comparison of root-mean-square deviation (RMSD) in geometry obtained using various methods. The shorthand notations for various methods are as follows: MINIs = HF-D3/MINIs, MINIs-ACP = HF-D3/MINIs-ACP, MINIX = HF-D3/MINIX, MINIX-ACP = HF-D3/MINIX-ACP, 6-31G* = HF-D3/6-31G*, and 6-31G*-ACP = HF-D3/6-31G*-ACP. All RMSDs are reported in Å unit.

Molecule	MINIs	MINIs-ACP	MINIX	MINIX-ACP	HF-3c	HF-3c-ACP	6-31G*	6-31G*-ACP
MOLdef_acetamide	0.499	0.025	0.499	0.031	0.031	0.030	0.019	0.012
MOLdef_acetic-acid	0.054	0.020	0.054	0.025	0.036	0.026	0.016	0.008
MOLdef_acetic-anhydride	0.214	0.192	0.214	0.204	0.276	0.203	0.043	0.043
MOLdef_acetone	0.033	0.015	0.033	0.020	0.026	0.020	0.017	0.008
MOLdef_acetylenamine	0.031	0.106	0.031	0.107	0.107	0.107	0.034	0.033

Molecule	MINIs	MINIs-ACP	MINIX	MINIX-ACP	HF-3c	HF-3c-ACP	6-31G*	6-31G*-ACP
MOLdef_acetylene	0.007	0.008	0.007	0.011	0.009	0.010	0.010	0.007
MOLdef_allene	0.014	0.015	0.014	0.016	0.006	0.016	0.012	0.008
MOLdef_aniline	0.025	0.093	0.025	0.092	0.079	0.092	0.029	0.022
MOLdef_azaborine	0.020	0.041	0.020	0.032	0.011	0.029	0.014	0.011
MOLdef_b2cl4	0.090	0.040	0.043	0.024	0.040	0.287	0.004	0.008
MOLdef_b2f4	0.069	0.024	0.069	0.027	0.051	0.028	0.007	0.014
MOLdef_b2h6	0.047	0.053	0.047	0.058	0.052	0.063	0.006	0.016
MOLdef_benzene	0.011	0.034	0.011	0.030	0.002	0.029	0.015	0.004
MOLdef_borane	0.002	0.009	0.002	0.000	0.002	0.003	0.002	0.002
MOLdef_borane_dms	0.109	0.076	0.058	0.037	0.077	0.035	0.039	0.045
MOLdef_borane_tbunh3	0.160	0.181	0.160	0.152	0.031	0.035	0.153	0.092
MOLdef_borane_thf	0.091	0.047	0.091	0.082	0.107	0.089	0.127	0.156
MOLdef_borazine	0.018	0.039	0.018	0.028	0.006	0.024	0.012	0.008
MOLdef_boricacid	0.061	0.010	0.061	0.016	0.052	0.023	0.022	0.028
MOLdef_c2h3f	0.031	0.030	0.031	0.029	0.018	0.044	0.014	0.013
MOLdef_c6f6	0.077	0.020	0.077	0.022	0.026	0.033	0.019	0.010
MOLdef_ccl4	0.086	0.020	0.026	0.017	0.019	0.228	0.024	0.004
MOLdef_cf4	0.078	0.021	0.078	0.008	0.018	0.010	0.019	0.006
MOLdef_ch2cl2	0.055	0.022	0.024	0.028	0.026	0.159	0.017	0.012
MOLdef_ch2f2	0.047	0.037	0.047	0.043	0.026	0.049	0.017	0.012
MOLdef_chlorobenzene	0.030	0.040	0.018	0.035	0.007	0.035	0.015	0.007
MOLdef_cis-butadiene	0.254	0.114	0.254	0.101	0.198	0.093	0.029	0.019
MOLdef_cl2	0.076	0.018	0.014	0.031	0.004	0.028	0.020	0.035
MOLdef_cl2o2s	0.483	0.024	0.037	0.042	0.024	0.492	0.045	0.030
MOLdef_cl2s2	0.123	0.083	0.081	0.080	0.067	0.545	0.087	0.077
MOLdef_clf3	0.103	0.087	0.043	0.097	0.104	0.059	0.059	0.031
MOLdef_clno	0.055	0.025	0.075	0.025	0.059	0.020	0.045	0.011
MOLdef_cs2	0.070	0.072	0.022	0.043	0.003	0.051	0.009	0.024
MOLdef_cumulene	0.018	0.025	0.018	0.027	0.006	0.026	0.014	0.011
MOLdef_cyclobutane	0.170	0.109	0.170	0.064	0.205	0.053	0.040	0.025
MOLdef_cyclopropane	0.016	0.010	0.016	0.011	0.013	0.011	0.013	0.003
MOLdef_diacetylene	0.028	0.021	0.028	0.021	0.011	0.021	0.010	0.021
MOLdef_diazene	0.029	0.037	0.029	0.044	0.025	0.049	0.017	0.014
MOLdef_dimethylether	0.036	0.026	0.036	0.028	0.024	0.028	0.026	0.018
MOLdef_ethane	0.008	0.014	0.008	0.015	0.007	0.016	0.012	0.005
MOLdef_ethenamine	0.025	0.036	0.025	0.038	0.042	0.041	0.052	0.052
MOLdef_ethenol	0.034	0.022	0.034	0.024	0.015	0.032	0.018	0.012
MOLdef_ethylene	0.005	0.012	0.005	0.013	0.006	0.013	0.011	0.002
MOLdef_ethynol	0.040	0.009	0.040	0.010	0.016	0.010	0.019	0.020
MOLdef_f2	0.009	0.059	0.009	0.076	0.067	0.081	0.022	0.001
MOLdef_f2o	0.033	0.024	0.033	0.028	0.049	0.027	0.032	0.006
MOLdef_fcl	0.039	0.049	0.011	0.032	0.039	0.012	0.024	0.012
MOLdef_foof	0.085	0.082	0.085	0.087	0.094	0.089	0.087	0.060

Molecule	MINIs	MINIs-ACP	MINIX	MINIX-ACP	HF-3c	HF-3c-ACP	6-31G*	6-31G*-ACP
MOLdef_formaldehyde	0.028	0.027	0.028	0.028	0.011	0.028	0.014	0.007
MOLdef_h2	0.020	0.029	0.020	0.023	0.020	0.021	0.008	0.026
MOLdef_h2o	0.020	0.012	0.020	0.012	0.011	0.011	0.009	0.016
MOLdef_hcl	0.032	0.055	0.012	0.040	0.020	0.045	0.009	0.015
MOLdef_hcn	0.017	0.011	0.017	0.012	0.005	0.010	0.009	0.004
MOLdef_hf	0.022	0.007	0.022	0.001	0.012	0.001	0.005	0.001
MOLdef_hno	0.038	0.024	0.038	0.027	0.011	0.030	0.018	0.009
MOLdef_hooh	0.254	0.067	0.254	0.069	0.044	0.122	0.036	0.043
MOLdef_hydrazine	0.022	0.016	0.022	0.026	0.029	0.032	0.031	0.045
MOLdef_hydroxylamine	0.030	0.012	0.030	0.018	0.025	0.019	0.019	0.021
MOLdef_ketene	0.036	0.016	0.036	0.014	0.009	0.012	0.010	0.008
MOLdef_methanimine	0.022	0.025	0.022	0.030	0.020	0.031	0.014	0.006
MOLdef_methylamine	0.014	0.018	0.014	0.019	0.014	0.019	0.022	0.023
MOLdef_methylsilane	0.017	0.033	0.013	0.027	0.012	0.030	0.009	0.021
MOLdef_n2	0.030	0.005	0.030	0.004	0.004	0.010	0.006	0.002
MOLdef_n3p3cl6	0.257	0.313	0.016	0.065	0.027	0.072	0.041	0.029
MOLdef_nf3	0.059	0.027	0.059	0.024	0.017	0.018	0.030	0.015
MOLdef_nitromethane	0.068	0.057	0.068	0.063	0.030	0.063	0.024	0.011
MOLdef_n-methylethenamine	0.134	0.101	0.134	0.096	0.097	0.092	0.055	0.031
MOLdef_n-methylmethanimine	0.028	0.024	0.028	0.027	0.027	0.027	0.019	0.009
MOLdef_o2	0.045	0.004	0.045	0.005	0.002	0.008	0.015	0.011
MOLdef_o3	0.085	0.030	0.085	0.030	0.025	0.037	0.027	0.022
MOLdef_ocs	0.076	0.019	0.042	0.018	0.010	0.023	0.008	0.007
MOLdef_p4s4	0.110	0.231	0.037	0.149	0.023	0.357	0.053	0.064
MOLdef_pf3	0.061	0.073	0.008	0.040	0.054	0.046	0.025	0.032
MOLdef_pf5	0.083	0.067	0.012	0.048	0.061	0.055	0.025	0.024
MOLdef_phenol	0.028	0.037	0.028	0.032	0.010	0.033	0.018	0.010
MOLdef_phenylacetylene	0.020	0.049	0.020	0.044	0.006	0.043	0.016	0.012
MOLdef_phenylboronicacid	0.031	0.055	0.031	0.042	0.030	0.040	0.021	0.019
MOLdef_propene	0.018	0.024	0.018	0.027	0.012	0.029	0.017	0.009
MOLdef_propyne	0.017	0.018	0.017	0.018	0.009	0.018	0.009	0.013
MOLdef_p-phenylenediamine	0.031	0.036	0.031	0.033	0.020	0.034	0.020	0.010
MOLdef_p-xylylene	0.018	0.048	0.018	0.044	0.011	0.043	0.020	0.008
MOLdef_quinone	0.040	0.021	0.040	0.015	0.024	0.014	0.017	0.009
MOLdef_s2	0.086	0.043	0.015	0.017	0.008	0.015	0.015	0.025
MOLdef_scl2	0.104	0.031	0.027	0.023	0.019	0.543	0.038	0.022
MOLdef_sf6	0.124	0.088	0.019	0.006	0.071	0.014	0.037	0.015
MOLdef_sh2	0.043	0.061	0.009	0.041	0.023	0.049	0.012	0.032
MOLdef_si2	0.001	0.057	0.039	0.004	0.039	0.003	0.002	0.031
MOLdef_si2h6	0.020	0.034	0.019	0.061	0.013	0.074	0.009	0.040
MOLdef_sicl2	0.097	0.047	0.018	0.102	0.012	0.135	0.022	0.028
MOLdef_sicl4	0.140	0.007	0.008	0.009	0.003	0.020	0.012	0.029
MOLdef_sif4	0.029	0.043	0.008	0.046	0.057	0.052	0.018	0.026

Molecule	MINIs	MINIs-ACP	MINIX	MINIX-ACP	HF-3c	HF-3c-ACP	6-31G*	6-31G*-ACP
MOLdef_sih4	0.012	0.024	0.017	0.036	0.003	0.044	0.007	0.033
MOLdef_sio	0.015	0.007	0.002	0.011	0.022	0.015	0.012	0.013
MOLdef_so	0.069	0.014	0.027	0.000	0.003	0.001	0.015	0.014
MOLdef_so2	0.137	0.030	0.033	0.033	0.022	0.032	0.024	0.027
MOLdef_so3	0.234	0.013	0.036	0.025	0.031	0.023	0.027	0.025
MOLdef_styrene	0.033	0.057	0.033	0.052	0.022	0.051	0.021	0.011
MOLdef_tetramethylsilane	0.025	0.075	0.008	0.041	0.015	0.035	0.013	0.017
MOLdef_thiophene	0.049	0.053	0.021	0.031	0.020	0.035	0.013	0.012
MOLdef_toluene	0.014	0.046	0.014	0.040	0.006	0.040	0.017	0.006
MOLdef_trans-butadiene	0.015	0.028	0.015	0.029	0.010	0.029	0.018	0.006
MOLdef_vinylacetylene	0.026	0.030	0.026	0.034	0.018	0.034	0.020	0.016
Riplinger_22-paracyclophane	0.069	0.094	0.069	0.089	0.158	0.101	0.129	0.104
Riplinger_23-dimethylbut-2-ene	0.088	0.184	0.088	0.144	0.146	0.180	0.029	0.119
Riplinger_2233-tetramethylbutane	0.036	0.058	0.036	0.048	0.019	0.053	0.060	0.042
Riplinger_C12H12_D6hcage	0.015	0.068	0.015	0.054	0.011	0.052	0.022	0.007
Riplinger_C12H12_cp-tropenyl	0.064	0.107	0.064	0.106	0.046	0.108	0.048	0.040
Riplinger_H_ttt	0.014	0.048	0.014	0.037	0.014	0.037	0.031	0.006
Riplinger_H_x+g-x+	0.079	0.051	0.079	0.047	0.040	0.051	0.056	0.037
Riplinger_Phe-Gly-Gly_elongated	0.262	0.113	0.262	0.110	0.092	0.124	0.145	0.085
Riplinger_Phe-Gly-Gly_folded	0.166	0.110	0.166	0.117	0.093	0.114	0.102	0.074
Riplinger_anthracene	0.082	0.111	0.082	0.097	0.057	0.099	0.037	0.042
Riplinger_diclophenac	0.150	0.145	0.171	0.154	0.117	0.143	0.132	0.100
Riplinger_dihydrofuran-2_3H-one	0.060	0.046	0.060	0.044	0.097	0.037	0.039	0.023
Riplinger_dimethylperoxide	0.061	0.030	0.061	0.033	0.035	0.032	0.032	0.017
Riplinger_ethanediol	0.111	0.038	0.111	0.043	0.065	0.044	0.033	0.021
Riplinger_heptahexane	0.036	0.050	0.036	0.054	0.006	0.053	0.018	0.022
Riplinger_heptatriyne	0.053	0.046	0.053	0.044	0.019	0.044	0.010	0.038
Riplinger_hexanoic_acid	0.078	0.055	0.078	0.048	0.047	0.048	0.045	0.020
Riplinger_methyl_pivalate	0.050	0.034	0.050	0.029	0.034	0.029	0.029	0.011
Riplinger_neo-pentane	0.009	0.035	0.009	0.027	0.010	0.027	0.022	0.003
Riplinger_n-octane	0.017	0.065	0.017	0.048	0.019	0.048	0.038	0.006
Riplinger_n-pentane	0.015	0.040	0.015	0.032	0.012	0.033	0.028	0.008
Riplinger_octamethylcyclobutane	0.034	0.072	0.034	0.057	0.023	0.056	0.036	0.014
Riplinger_penicilline	0.555	0.513	0.538	0.513	0.531	0.620	0.331	0.135
Riplinger_pentan-24-dione	0.184	0.101	0.184	0.083	0.193	0.074	0.039	0.016
Riplinger_p-xylene	0.014	0.058	0.014	0.051	0.006	0.050	0.020	0.007
Riplinger_tetrahydropyran-2-one	0.109	0.053	0.109	0.051	0.062	0.049	0.045	0.039
Riplinger_vancomycine	2.230	1.273	0.430	1.276	1.358	1.286	0.311	0.154
Riplinger_vinylacetate	0.078	0.027	0.078	0.032	0.032	0.044	0.024	0.010
Hairpin_C14H30_folded	0.157	0.067	0.157	0.060	0.054	0.063	0.104	0.015
Hairpin_C14H30_linear	0.037	0.104	0.037	0.073	0.042	0.072	0.050	0.013
Hairpin_C15H32_folded	0.184	0.067	0.184	0.062	0.061	0.061	0.113	0.025
Hairpin_C15H32_linear	0.040	0.111	0.040	0.078	0.045	0.077	0.053	0.013

Molecule	MINIs	MINIs-ACP	MINIX	MINIX-ACP	HF-3c	HF-3c-ACP	6-31G*	6-31G*-ACP
Hairpin_C16H34_folded	0.152	0.070	0.152	0.061	0.043	0.066	0.085	0.017
Hairpin_C16H34_linear	0.043	0.118	0.043	0.083	0.048	0.082	0.056	0.014
Hairpin_C17H36_folded	0.153	0.077	0.153	0.069	0.051	0.072	0.087	0.015
Hairpin_C17H36_linear	0.046	0.126	0.046	0.089	0.050	0.087	0.059	0.015
Hairpin_C18H38_folded	0.155	0.083	0.155	0.073	0.056	0.076	0.091	0.016
Hairpin_C18H38_linear	0.048	0.133	0.048	0.094	0.053	0.092	0.063	0.016
Peptides_1yjp	0.501	1.222	0.501	1.236	0.568	1.217	0.237	0.160
Peptides_2omm	0.493	0.742	0.493	0.763	0.582	0.738	0.205	0.159
Peptides_2y29	0.980	0.372	0.980	0.273	0.996	0.298	0.651	0.481
Peptides_3dg1	0.421	0.254	0.421	0.276	0.306	0.276	0.271	0.151
Peptides_3fpo	0.695	0.573	0.695	0.565	0.554	1.933	0.390	0.210
Peptides_3ftk	1.144	0.758	1.144	0.943	0.318	0.964	0.236	0.225
Peptides_3ftr	0.378	0.284	0.378	0.277	0.352	0.264	0.279	0.158
Peptides_3fva	0.829	1.159	0.829	1.174	1.107	0.548	0.546	0.109
Peptides_3nvg	0.551	0.407	0.573	0.403	0.557	0.354	0.201	0.242
Peptides_3q2x	0.684	3.382	0.684	3.379	1.886	3.286	1.042	0.405
Peptides_3sgs	3.940	0.549	3.940	0.537	0.581	0.972	2.894	0.467
Peptides_4nip	0.872	0.370	0.872	0.397	0.646	0.458	0.188	0.126
Peptides_4qxx	3.767	0.786	3.767	0.862	3.638	0.845	0.375	0.336
Peptides_4r0u	0.791	2.181	0.791	2.007	2.305	2.042	0.202	0.188
Peptides_4r0w	1.400	0.416	1.400	0.468	0.672	0.480	0.285	0.193
LB12_BHS	0.381	0.264	0.071	0.165	0.047	0.175	0.072	0.061
LB12_DIAD	0.048	0.122	0.048	0.095	0.024	0.092	0.056	0.021
LB12_DTFS	0.060	0.094	0.280	0.129	0.272	0.149	0.066	0.049
LB12_FLP	0.246	0.247	0.240	0.232	0.186	0.276	0.129	0.116
LB12_MESITRAN	0.068	0.106	0.285	0.109	0.187	0.083	0.053	0.041
LB12_PP	0.078	0.113	0.071	0.087	0.041	0.087	0.047	0.050
LB12_RESVAN	0.092	0.102	0.087	0.086	0.095	0.095	0.038	0.029
LB12_S82+	0.277	0.214	0.141	0.201	0.141	0.797	0.148	0.190
CLB18_HOLKEY01	0.280	0.160	0.288	0.187	0.278	0.207	0.059	0.036
CLB18_HOLKOI01	0.067	0.115	0.068	0.098	0.058	0.104	0.063	0.056
CLB18_JOWROF	0.270	0.465	0.270	0.488	0.481	0.543	0.095	0.074
CLB18_RIRTUH	0.039	0.106	0.039	0.090	0.015	0.095	0.030	0.037
CLB18_UBEQAV	0.048	0.113	0.048	0.088	0.025	0.085	0.053	0.020
CLB18_YISRUQ	0.264	0.319	0.264	0.401	0.117	0.127	0.048	0.052
CLB18_YISSAX	0.067	0.072	0.067	0.163	0.178	0.191	0.063	0.020
CLB18_YISSEB	0.544	0.449	0.544	0.464	0.422	0.508	0.178	0.217
CLB18_YISSIF	2.539	2.320	2.539	2.432	2.266	2.430	0.154	0.109
CLB18_YISSOL	0.404	1.118	0.404	1.176	0.435	1.165	0.157	0.135
CLB18_dihydropyracilene_10a	0.097	0.122	0.097	0.108	0.071	0.116	0.073	0.071
CLB18_dihydropyracilene_10b	0.093	0.129	0.093	0.113	0.071	0.121	0.070	0.069
CLB18_dihydropyracilene_10c	0.097	0.127	0.097	0.112	0.074	0.120	0.073	0.072
A21_01_water-ammonia	0.107	0.097	0.107	0.095	0.095	0.095	0.031	0.057

Molecule	MINIs	MINIs-ACP	MINIX	MINIX-ACP	HF-3c	HF-3c-ACP	6-31G*	6-31G*-ACP
A21_02_water-dimer	0.069	0.055	0.069	0.056	0.075	0.058	0.104	0.082
A21_03_HCN-dimer	0.038	0.014	0.038	0.029	0.021	0.031	0.028	0.017
A21_04_HF-dimer	0.062	0.055	0.062	0.059	0.043	0.051	0.463	0.163
A21_05_ammonia-dimer	0.056	0.018	0.056	0.020	0.034	0.017	0.051	0.040
A21_06_HF-methane	0.031	0.044	0.031	0.044	0.136	0.070	0.029	0.013
A21_07_ammonia-methane	0.246	0.127	0.246	0.128	0.186	0.140	0.037	0.073
A21_08_water-methane	0.455	0.279	0.455	0.264	0.055	0.021	0.805	0.459
A21_09_formaldehyde-dimer	0.052	0.058	0.052	0.063	0.070	0.070	0.038	0.051
A21_10_water-ethene	0.086	0.088	0.086	0.096	0.125	0.099	0.293	0.077
A21_11_formaldehyde-ethene	0.089	0.582	0.089	0.581	0.088	0.144	0.111	0.018
A21_12_ethyne-dimer	0.070	0.024	0.070	0.044	0.018	0.036	0.025	0.028
A21_13_ammonia-ethene	0.174	0.195	0.174	0.202	0.155	0.210	0.053	0.178
A21_14_ethene-dimer	0.141	0.015	0.141	0.017	0.048	0.018	0.021	0.041
A21_15_methane-ethene	0.229	0.469	0.229	0.469	0.237	0.462	0.458	0.462
A21_16_borane-methane	0.064	0.065	0.064	0.080	0.199	0.095	0.089	0.018
A21_17_methane-ethane	0.125	0.042	0.125	0.052	0.020	0.047	0.082	0.022
A21_18_methane-ethane	0.125	0.032	0.125	0.043	0.028	0.046	0.097	0.030
A21_19_methane-dimer	0.130	0.044	0.130	0.054	0.029	0.063	0.097	0.034
S66_AcNH2AcNH2	0.508	0.506	0.508	0.507	0.589	0.044	0.022	0.021
S66_AcNH2Uracil	0.365	0.363	0.365	0.364	0.358	0.045	0.034	0.029
S66_AcOHAcOH	0.063	0.043	0.063	0.045	0.063	0.047	0.027	0.017
S66_AcOHUracil	0.062	0.027	0.062	0.027	0.040	0.028	0.025	0.016
S66_BenzeneAcNH2NHpi	0.326	0.833	0.326	0.727	0.696	0.682	0.225	0.151
S66_BenzeneAcOH	0.214	0.168	0.214	0.170	0.085	0.654	0.246	0.069
S66_BenzeneAcOHOHpi	0.139	0.074	0.139	0.066	0.094	0.065	0.295	0.277
S66_BenzeneBenzeneTS	0.192	0.176	0.190	0.175	0.097	0.089	0.163	0.595
S66_BenzeneBenzenepipi	0.324	0.269	0.309	0.215	0.104	0.083	0.084	0.053
S66_BenzeneCyclopentane	0.160	0.083	0.160	0.085	0.084	0.097	0.074	0.042
S66_BenzeneEthene	0.349	1.012	0.349	1.011	0.066	0.054	0.735	0.739
S66_BenzeneEthyneCHpi	0.110	0.048	0.110	0.053	0.033	0.053	0.079	0.046
S66_BenzeneMeNH2NHpi	0.285	0.314	0.285	0.266	0.062	0.048	0.357	0.319
S66_BenzeneMeOHOHpi	0.324	0.068	0.324	0.056	0.094	0.053	0.108	0.083
S66_BenzeneNeopentane	0.165	0.112	0.165	0.120	0.080	0.102	0.069	0.029
S66_BenzenePeptideNHpi	0.527	0.574	0.527	0.613	0.356	0.161	0.651	0.266
S66_BenzenePyridineTS	0.289	0.082	0.288	0.082	0.049	0.061	0.105	0.614
S66_BenzenePyridinepipi	0.260	0.193	0.260	0.200	0.093	0.113	0.171	0.228
S66_BenzeneUracilpipi	0.537	0.221	0.537	0.225	0.069	0.300	0.339	0.559
S66_BenzeneWaterOHpi	0.235	0.216	0.235	0.194	0.130	0.130	0.253	0.227
S66_CyclopentaneCyclopentane	0.195	0.097	0.195	0.082	0.113	0.071	0.068	0.019
S66_CyclopentaneNeopentane	0.203	0.082	0.203	0.090	0.078	0.055	0.483	0.186
S66_EthenePentane	0.192	0.249	0.192	0.242	0.037	0.035	0.080	0.048
S66_EthyneAcOHOHpi	0.103	0.058	0.103	0.068	0.063	0.066	0.077	0.021
S66_EthyneEthyneTS	0.108	0.061	0.108	0.082	0.049	0.073	0.060	0.012

Molecule	MINIs	MINIs-ACP	MINIX	MINIX-ACP	HF-3c	HF-3c-ACP	6-31G*	6-31G*-ACP
S66_EthynePentane	0.207	0.034	0.207	0.027	0.053	0.025	0.212	0.150
S66_EthyneWaterCHO	0.172	0.092	0.172	0.095	0.069	0.055	0.281	0.300
S66_MeNH2MeNH2	0.243	0.235	0.243	0.261	0.225	0.186	0.076	0.052
S66_MeNH2MeOH	0.515	1.369	0.515	1.336	0.243	0.364	1.423	1.384
S66_MeNH2Peptide	0.784	0.439	0.784	0.431	0.867	0.457	0.331	0.289
S66_MeNH2Pyridine	0.227	0.879	0.227	0.884	0.761	0.787	0.808	0.887
S66_MeNH2Water	0.124	0.168	0.125	0.170	0.176	0.171	0.307	0.255
S66_MeOHMeNH2	0.253	0.265	0.253	0.242	0.037	0.032	0.632	0.741
S66_MeOHMeOH	0.304	0.130	0.304	0.133	0.137	0.125	0.426	0.296
S66_MeOHPeptide	0.396	0.582	0.396	0.919	0.456	0.608	0.234	0.149
S66_MeOHPyridine	0.476	0.062	0.476	0.053	0.040	0.050	0.559	0.443
S66_MeOHWater	0.068	0.067	0.068	0.059	0.033	0.034	0.117	0.127
S66_NeopentaneNeopentane	0.201	0.100	0.201	0.092	0.093	0.081	0.079	0.011
S66_NeopentanePentane	0.169	0.084	0.169	0.086	0.047	0.039	0.077	0.023
S66_PentaneAcNH2	0.349	0.268	0.349	0.271	0.402	0.160	0.150	0.070
S66_PentaneAcOH	0.309	0.327	0.309	0.322	0.267	0.144	0.107	0.108
S66_PentanePentane	0.150	0.047	0.150	0.057	0.034	0.027	0.082	0.010
S66_PeptideEthene	0.272	0.201	0.272	0.198	0.242	0.201	0.415	0.351
S66_PeptideMeNH2	0.559	0.524	0.559	0.524	0.512	0.274	0.174	0.055
S66_PeptideMeOH	0.426	0.468	0.426	0.468	0.555	0.370	0.216	0.180
S66_PeptidePentane	0.232	0.178	0.232	0.155	0.199	0.159	0.095	0.055
S66_PeptidePeptide	0.588	0.489	0.588	0.471	0.559	0.507	0.429	0.428
S66_PeptideWater	0.274	0.186	0.274	0.166	0.428	0.076	0.466	0.473
S66_PyridineEthene	0.144	1.042	0.144	1.044	0.054	0.283	0.947	1.042
S66_PyridineEthyne	0.623	0.083	0.623	0.096	0.068	0.089	0.733	0.690
S66_PyridinePyridineCHN	0.123	0.094	0.123	0.098	0.069	0.101	0.072	0.019
S66_PyridinePyridineTS	0.129	0.075	0.129	0.079	0.056	0.074	0.183	0.330
S66_PyridinePyridinepipi	0.125	0.089	0.125	0.090	0.074	0.102	0.228	0.072
S66_PyridineUracilpipi	0.145	0.185	0.145	0.201	0.065	0.241	0.165	0.061
S66_UracilCyclopentane	0.229	0.080	0.229	0.088	0.088	0.082	0.116	0.080
S66_UracilEthene	0.128	0.092	0.128	0.084	0.100	0.154	0.116	0.053
S66_UracilEthyne	0.147	0.045	0.147	0.040	0.094	0.052	0.128	0.092
S66_UracilNeopentane	0.191	0.165	0.191	0.153	0.078	0.146	0.109	0.057
S66_UracilPentane	0.166	0.061	0.166	0.061	0.063	0.075	0.083	0.033
S66_UracilUracilBP	0.073	0.045	0.073	0.039	0.025	0.039	0.026	0.015
S66_UracilUracilpipi	0.094	0.071	0.094	0.096	0.138	0.094	0.146	0.142
S66_WaterMeNH2	0.427	0.121	0.427	0.110	0.123	0.110	0.390	0.348
S66_WaterMeOH	0.075	0.155	0.075	0.155	0.183	0.149	0.252	0.180
S66_WaterPeptide	0.437	0.414	0.437	0.052	0.070	0.054	0.426	0.439
S66_WaterPyridine	0.054	0.074	0.054	0.070	0.039	0.068	0.413	0.364
S66_WaterWater	0.073	0.046	0.073	0.044	0.058	0.034	0.129	0.105
L7_c2c2pd	0.164	0.128	0.164	0.120	0.113	0.120	0.074	0.068
L7_c3a	0.481	0.484	0.481	0.479	0.326	0.857	0.416	0.407

Molecule	MINIs	MINIs-ACP	MINIX	MINIX-ACP	HF-3c	HF-3c-ACP	6-31G*	6-31G*-ACP
L7_c3gc	0.199	0.227	0.199	0.215	0.156	0.201	0.174	0.154
L7_cbh	0.678	0.638	0.678	0.696	0.061	0.631	0.724	0.629
L7_gcgc	0.637	0.657	0.637	0.673	0.648	0.703	0.567	0.563
L7_ggg	3.377	3.604	3.377	3.598	1.937	1.216	3.366	2.440
L7_phe	2.466	2.000	2.466	1.981	2.093	1.870	2.400	0.192
S30L_01_complex	0.275	0.155	0.275	0.136	0.071	0.148	0.100	0.050
S30L_02_complex	0.269	0.143	0.269	0.124	0.124	0.131	0.153	0.139
S30L_03_complex	0.542	0.514	0.542	0.510	0.483	0.301	0.224	0.163
S30L_04_complex	0.259	0.170	0.264	0.152	0.172	0.156	0.205	0.136
S30L_05_complex	0.817	0.118	0.816	0.111	0.600	0.135	0.118	0.072
S30L_06_complex	0.310	0.438	0.310	0.406	0.332	0.400	0.375	0.420
S30L_07_complex	0.330	0.706	0.330	0.700	0.223	0.258	0.922	0.161
S30L_08_complex	0.787	0.790	0.787	0.780	0.267	0.729	0.724	0.753
S30L_09_complex	0.085	0.186	0.085	0.161	0.042	0.173	0.047	0.027
S30L_10_complex	0.144	0.278	0.144	0.243	0.088	0.248	0.075	0.078
S30L_11_complex	0.179	0.448	0.163	0.187	0.101	0.192	0.092	0.104
S30L_12_complex	0.203	0.226	0.187	0.256	0.125	0.244	0.887	0.554
S30L_13_complex	0.455	0.268	0.455	0.179	0.167	0.565	0.255	0.111
S30L_14_complex	0.469	0.202	0.476	0.177	0.164	0.553	0.254	0.115
S30L_17_complex	0.367	0.354	0.367	0.357	0.416	0.399	0.126	0.101
S30L_18_complex	0.204	0.722	0.204	0.904	0.176	0.243	0.528	0.545
S30L_19_complex	0.397	0.194	0.397	0.183	0.108	0.188	0.384	0.161
S30L_20_complex	0.300	0.248	0.300	0.233	0.112	0.241	0.334	0.220
S30L_21_complex	0.373	0.156	0.373	0.153	0.110	0.166	1.176	0.129
S30L_22_complex	1.645	0.291	1.645	0.356	2.330	0.428	1.697	0.558
S30L_23_complex	0.666	0.671	0.666	0.666	0.653	0.666	0.359	0.219
S30L_24_complex	0.091	0.192	0.091	0.115	0.050	0.122	0.204	0.168
S30L_25_complex	0.501	0.247	0.501	0.229	0.156	0.233	0.161	0.080
S30L_26_complex	0.507	0.242	0.507	0.225	0.165	0.227	0.142	0.070
S30L_27_complex	0.130	0.118	0.130	0.115	0.065	0.125	0.207	0.067
S30L_28_complex	0.213	0.152	0.213	0.151	0.297	0.218	0.329	0.107
S30L_29_complex	0.501	0.162	0.501	0.147	0.232	0.159	0.301	0.113
S30L_30_complex	0.887	0.177	0.885	0.172	0.431	0.184	0.200	0.089

Table S6. Comparison of percentage change in the single-point (SP) calculation time between that of uncorrected and ACP corrected approaches for selected molecules. The shorthand notations used are as follows: MINIs = [((SP time of HF-D3/MINIs-ACP) – (SP time of HF-D3/MINIs)) / (SP time of HF-D3/MINIs) x 100%], MINIX = [((SP time of HF-D3/MINIX-ACP) – (SP time of HF-D3/MINIX)) / (SP time of HF-D3/MINIX) x 100%], and 6-31G* = [((SP time of HF-D3/6-31G*-ACP) – (SP time of HF-D3/6-31G*)) / (SP time of HF-D3/MINIs) x 100%]. Single-point calculations were performed using Gaussian16 package and 32 cores of Dell EMC R440 CPU compute nodes on *Sockeye* cluster (University of British Columbia's Advanced Research Computing facility).

Molecule	**Atoms**	**MINIs (in %)**	**MINIX (in %)**	**6-31G* (in %)**
Dichlophenac	30	12.5	13.5	9.7
Penicilline	42	19.0	19.6	7.3
Anthracene	48	11.1	20.0	0.0
a_UDPy complex	126	20.8	21.6	3.1
c_dendrimer complex	144	31.7	23.0	-1.7
d_Cyc-Pep complex	160	27.9	25.4	2.2
e_AB3-Peptide complex	174	31.1	24.6	-1.5
Vancomycine	176	26.2	13.7	3.8
1_capsule complex	200	19.9	18.3	5.2
2_Cyc-Pep-2 complex	296	28.6	15.8	-2.4
4_ALA-BN complex	381	20.3	21.5	-7.3
$C_{150}N_{50}O_{51}H_{252}$	503	15.4	9.0	-0.3
5_gramicidin complex	552	21.4	3.6	8.3
Crambin*	644	76.9	-33.3	-42.1
6_helix-rod complex	750	30.3	3.8	9.1
Insulin*	787	28.2	10.0	93.0
7_DNA complex	910	41.0	11.3	5.0
8_Protein-Ligand complex	1027	31.4	6.0	-1.8
$C_{350}H_{702}$	1052	7.1	18.3	-10.1
Integrase	2380	16.4	10.2	4.2

* For the case of Crambin, the number of SCF cycles without and with ACP were respectively, 43 and 54 (for HF-D3/MINIS), 57 and 51 (for HF-D3/MINIX), and 65 and 45 (for HF-D3/6-31G*). Whereas, for the case of Insulin, the number of SCF cycles without and with ACP for the HF-D3/6-31G* method was 19 and 26, respectively.

Appendix 6

Supporting Information for Chapter 9

Section S1. Sample input file demonstrating the use of atom-centered potentials in Gaussian16 software

The 6-31G* basis set file (in .gbs extension) and the corresponding ACP files (in .acp extension) are provided separately in the supporting information ZIP file accompanying this document. An externally specified basis set file named **“631gs.gbs”** and the additional ACP file **“631gs.acp”** is defined and invoked by adding the keyword **“genECP”** to the route section of the Gaussian input file. Note that the ACP are not transferable and are proposed to be used with their underlying methods only.

```
%mem=4GB
%nprocs=8
# BLYP empiricaldispersion=gd3bj genECP

Title: Sample water dimer input using BLYP-D3/6-31G*-ACP method

0 1
O   -0.702196054   -0.056060256    0.009942262
H   -1.022193224    0.846775782   -0.011488714
H    0.257521062    0.042121496    0.005218999
O    2.220871067    0.026716792    0.000620476
H    2.597492682   -0.411663274    0.766744858
H    2.593135384   -0.449496183   -0.744782026

@631gs.gbs/N

@631gs.acp/N
```

Section S2. 6-31G* basis set file

```
-H     0
S   3   1.00
     18.7311370            0.03349460
      2.8253937            0.23472695
      0.6401217            0.81375733
S   1   1.00
      0.1612778            1.0000000
****
-B     0
S   6   1.00
   2068.8823000            0.0018663
    310.6495700            0.0142515
     70.6830330            0.0695516
     19.8610800            0.2325729
      6.2993048            0.4670787
      2.1270270            0.3634314
SP   3   1.00
      4.7279710           -0.1303938            0.0745976
      1.1903377           -0.1307889            0.3078467
      0.3594117            1.1309444            0.7434568
SP   1   1.00
      0.1267512            1.0000000            1.0000000
```

```
D   1   1.00
      0.6000000              1.0000000
****
-C     0
S   6   1.00
   3047.5249000              0.0018347
    457.3695100              0.0140373
    103.9486900              0.0688426
     29.2101550              0.2321844
      9.2866630              0.4679413
      3.1639270              0.3623120
SP   3   1.00
      7.8682724             -0.1193324              0.0689991
      1.8812885             -0.1608542              0.3164240
      0.5442493              1.1434564              0.7443083
SP   1   1.00
      0.1687144              1.0000000              1.0000000
D   1   1.00
      0.8000000              1.0000000
****
-N     0
S   6   1.00
   4173.5110000              0.0018348
    627.4579000              0.0139950
    142.9021000              0.0685870
     40.2343300              0.2322410
     12.8202100              0.4690700
      4.3904370              0.3604550
SP   3   1.00
     11.6263580             -0.1149610              0.0675800
      2.7162800             -0.1691180              0.3239070
      0.7722180              1.1458520              0.7408950
SP   1   1.00
      0.2120313              1.0000000              1.0000000
D   1   1.00
      0.8000000              1.0000000
****
-O     0
S   6   1.00
   5484.6717000              0.0018311
    825.2349500              0.0139501
    188.0469600              0.0684451
     52.9645000              0.2327143
     16.8975700              0.4701930
      5.7996353              0.3585209
SP   3   1.00
     15.5396160             -0.1107775              0.0708743
      3.5999336             -0.1480263              0.3397528
      1.0137618              1.1307670              0.7271586
SP   1   1.00
      0.2700058              1.0000000              1.0000000
D   1   1.00
      0.8000000              1.0000000
****
-F     0
S   6   1.00
   7001.7130900              0.0018196169
   1051.3660900              0.0139160796
    239.2856900              0.0684053245
     67.3974453              0.233185760
     21.5199573              0.471267439
      7.40310130             0.356618546
SP   3   1.00
     20.8479528             -0.108506975            0.0716287243
     4.80830834          -0.146451658          0.3459121030
     1.34406986           1.128688580          0.7224699570
SP   1   1.00
     0.358151393          1.0000000            1.0000000
D   1   1.00
     0.8000000           1.0000000
****
-Si     0
S   6   1.00
  16115.9000000           0.00195948
   2425.5800000           0.01492880
    553.8670000           0.07284780
    156.3400000           0.24613000
     50.0683000          0.48591400
     17.0178000          0.32500200
SP   6   1.00
    292.7180000           -0.00278094           0.00443826
     69.8731000          -0.03571460           0.03266790
     22.3363000          -0.11498500           0.13472100
     8.1503900           0.09356340           0.32867800
     3.1345800           0.60301700           0.44964000
     1.2254300           0.41895900           0.26137200
SP   3   1.00
     1.7273800          -0.24463000          -0.01779510
     0.5729220           0.00431572           0.25353900
     0.2221920           1.09818000           0.80066900
SP   1   1.00
     0.0778369           1.00000000           1.00000000
D   1   1.00
     0.4500000           1.0000000
****
-P     0
S   6   1.00
  19413.3000000           0.0018516
   2909.4200000           0.0142062
    661.3640000           0.0699995
    185.7590000           0.2400790
     59.1943000          0.4847620
     20.0310000          0.3352000
SP   6   1.00
    339.4780000           -0.00278217           0.00456462
     81.0101000          -0.0360499           0.03369360
     25.8780000          -0.1166310           0.13975500
     9.4522100           0.0968328           0.33936200
     3.6656600           0.6144180           0.45092100
     1.4674600           0.4037980           0.23858600
SP   3   1.00
     2.1562300          -0.2529230          -0.01776530
     0.7489970           0.0328517           0.27405800
     0.2831450           1.0812500           0.78542100
SP   1   1.00
     0.0998317           1.0000000           1.00000000
D   1   1.00
     0.5500000           1.0000000
****
-S     0
S   6   1.00
  21917.1000000           0.0018690
   3301.4900000           0.0142300
    754.1460000           0.0696960
    212.7110000           0.2384870
     67.9896000          0.4833070
     23.0515000          0.3380740
SP   6   1.00
```

```
    423.7350000           -0.0023767            0.0040610
    100.7100000           -0.0316930            0.0306810
     32.1599000           -0.1133170            0.1304520
     11.8079000            0.0560900            0.3272050
      4.6311000            0.5922550            0.4528510
      1.8702500            0.4550060            0.2560420
SP   3   1.00
      2.6158400           -0.2503740           -0.0145110
      0.9221670            0.0669570            0.3102630
      0.3412870            1.0545100            0.7544830
SP   1   1.00
      0.1171670            1.0000000            1.0000000
D   1   1.00
      0.6500000            1.0000000
****
-Cl     0
S   6   1.00
  25180.1000000            0.0018330
   3780.3500000            0.0140340
    860.4740000            0.0690970
    242.1450000            0.2374520
     77.3349000            0.4830340
     26.2470000            0.3398560
SP   6   1.00
    491.7650000           -0.0022974            0.0039894
    116.9840000           -0.0307140            0.0303180
     37.4153000           -0.1125280            0.1298800
     13.7834000            0.0450160            0.3279510
      5.4521500            0.5893530            0.4535270
      2.2258800            0.4652060            0.2521540
SP   3   1.00
      3.1864900           -0.2518300           -0.0142990
      1.1442700            0.0615890            0.3235720
      0.4203770            1.0601800            0.7435070
SP   1   1.00
      0.1426570            1.0000000            1.0000000
D   1   1.00
      0.7500000            1.0000000
****
```

Section S3. ACP file for BLYP-D3/6-31G*

```
-H  0
H  1 0
l
7
2 0.120000 0.020147140373755
2 0.140000 -0.041237723349528
2 0.160000 -0.003307405765286
2 0.240000 0.059674586582327
2 0.400000 -0.118823903168496
2 0.700000 0.247290160402391
2 1.500000 -0.527150318071152
s
5
2 0.160000 0.120691251097858
2 0.200000 0.038930258383712
2 0.400000 -0.265355031436539
2 1.500000 0.274814946103966
2 2.500000 0.161357217122116
-B  0
B  3 0
l
```

```
2
2 0.140000 -0.005361095419549
2 0.160000 -0.003503927184227
s
1
2 0.120000 -0.031615803607151
p
2
2 0.120000 0.020325629999813
2 0.300000 -0.001350604360460
d
1
2 0.120000 0.007416401481579
-C  0
C  3 0
l
7
2 0.120000 0.014400120001587
2 0.140000 -0.052549979917686
2 0.200000 0.135079603515878
2 0.240000 -0.166403170082333
2 0.400000 0.062322713365957
2 0.600000 0.065877688424473
2 1.300000 -0.509124098696679
s
4
2 0.120000 -0.193746735235504
2 0.180000 0.309902121431847
2 0.280000 0.207867260822578
2 3.000000 -1.659236605095590
p
3
2 0.140000 0.076665306805187
2 0.300000 -0.226281377613320
2 0.800000 0.499350009929005
d
3
2 0.160000 0.121369724270459
2 0.400000 -0.230529028625699
2 1.000000 0.056682374294509
-N  0
N  3 0
l
5
2 0.120000 -0.012965150835205
2 0.160000 0.007256219105311
2 0.260000 -0.064977617385447
2 0.400000 0.103308634506769
2 0.900000 -0.069459655901241
s
1
2 0.120000 0.017404676193864
p
3
2 0.120000 0.062349439921297
2 0.260000 -0.071759517184964
2 1.200000 0.072764670657726
d
2
2 0.120000 0.094322328568004
2 0.500000 -0.297478725370155
-O  0
O  3 0
l
```

```
5
2 0.120000 -0.012871157707849
2 0.140000 0.014997546243208
2 0.220000 -0.076452164196361
2 0.300000 0.087768562473209
2 0.700000 -0.021187476231135
s
3
2 0.260000 -0.100373717477192
2 0.800000 0.582540412119212
2 2.500000 0.851963600682503
p
5
2 0.120000 0.147497238278610
2 0.180000 -0.104631787672340
2 0.500000 -0.192201180030414
2 1.500000 0.451726383445369
2 2.000000 0.244005547073623
d
4
2 0.120000 0.042314620277864
2 0.240000 0.147057122125556
2 0.500000 -0.513625296963165
2 1.500000 -0.497885246620486
-F  0
F  3 0
l
4
2 0.120000 -0.003083680165391
2 0.140000 -0.007250527071461
2 0.300000 -0.004042194634722
2 2.500000 0.355542061698070
s
1
2 0.140000 0.134874716784963
p
3
2 0.120000 0.156210558646745
2 0.260000 -0.072581580290277
2 0.600000 -0.280630251062100
d
1
2 0.160000 -0.000452996272237
-Si 0
Si 3 0
l
4
2 0.120000 -0.012768588855814
2 0.140000 -0.002982953967765
2 0.240000 0.000111132618359
2 1.600000 -0.000002073444067
s
3
2 0.120000 0.067007805867758
2 0.140000 -0.000003402021170
2 0.180000 0.000010280680605
p
1
2 0.120000 0.036514135099418
d
2
2 0.120000 0.056325194328794
2 1.800000 -0.000047090454828
-P  0
```

```
P  3 0
l
4
2 0.120000 -0.016063449165700
2 0.200000 0.011785314847653
2 0.400000 -0.012948602971997
2 0.500000 -0.092779692435669
s
2
2 0.120000 0.070158930317678
2 1.900000 -0.000005938475577
p
5
2 0.120000 0.057613896982525
2 0.220000 -0.000011099697332
2 0.500000 0.000004308530064
2 0.700000 0.000019322472574
2 3.000000 -0.000003616229676
d
5
2 0.120000 0.050471450279694
2 0.800000 0.000007865370229
2 0.900000 0.000012705757120
2 1.300000 0.000001331941009
2 1.600000 0.000013329588199
-S  0
S  3 0
l
9
2 0.120000 -0.027059548423689
2 0.140000 -0.001413696385708
2 0.220000 0.026328334660683
2 0.240000 0.000000957996311
2 0.500000 -0.070550491180111
2 0.700000 0.000003932750072
2 1.100000 -0.000012048772210
2 1.400000 -0.000001097199151
2 2.000000 -0.000020396609107
s
7
2 0.120000 0.000010559621141
2 0.140000 0.090413590939423
2 0.280000 -0.000001650984264
2 0.500000 -0.000016538374873
2 0.800000 -0.000013011986423
2 0.900000 -0.000004416127094
2 1.900000 0.000006748061738
p
9
2 0.120000 0.109819100072531
2 0.140000 -0.000000412421364
2 0.160000 0.000000123726788
2 0.200000 -0.000007396473012
2 0.260000 -0.073312863113150
2 0.280000 0.000001443410866
2 0.900000 -0.000007417834592
2 1.500000 0.000004288679994
2 1.700000 0.000001830691959
d
9
2 0.120000 0.162907088374595
2 0.240000 -0.112047009373014
2 0.400000 -0.263505507305333
2 0.500000 -0.000000490731182
```

```
2 0.600000 -0.000001084488249
2 0.800000 0.000003435286170
2 1.600000 -0.000001767773126
2 2.000000 -0.000013202051264
2 3.000000 0.000006119040848
-Cl 0
Cl 3 0
l
5
2 0.120000 -0.007839046988501
2 0.140000 -0.009904704409079
2 0.280000 0.000001092910593
2 0.400000 -0.042600256011967
2 1.100000 0.000028610404322
s
1
2 0.120000 0.232370656326083
p
2
2 0.120000 0.078480045097496
2 3.000000 0.000007282252064
d
3
2 0.120000 0.000015968835170
2 0.140000 0.050072455646507
2 1.200000 0.000002772909011
```

Section S4. ACP file for M062X/6-31G*

```
-H  0
H  1 0
l
6
2 0.120000 -0.000349949245644
2 0.260000 0.017203579390844
2 0.300000 -0.032154667302613
2 0.500000 0.053140470603465
2 0.800000 -0.053747258133887
2 1.000000 -0.033403317030972
s
4
2 0.120000 0.046171973387597
2 0.220000 -0.079798170492764
2 1.300000 0.059653512711890
2 2.000000 0.065309724938451
-B  0
B  3 0
l
3
2 0.120000 0.001933684886241
2 0.160000 -0.002856300840556
2 0.300000 0.006182729562557
s
1
2 0.180000 -0.011793261730715
p
2
2 0.120000 0.011768097404168
2 0.700000 -0.057231361601408
d
2
2 0.120000 -0.016650603763724
2 0.500000 0.035513229186388
```

```
-C  0
C  3 0
l
5
2 0.120000 0.002376355217199
2 0.180000 -0.005354604201182
2 0.280000 0.023427856436081
2 0.400000 -0.035476312175266
2 3.000000 -0.146163309149977
s
5
2 0.120000 -0.165071028585521
2 0.160000 0.212984002714891
2 0.200000 0.126314155692729
2 0.800000 -0.273239817939952
2 1.400000 -0.188213294259154
p
3
2 0.180000 -0.137009621201451
2 0.220000 0.113726980268623
2 0.500000 0.092392247514364
d
3
2 0.140000 0.016612170886135
2 0.260000 -0.080227866631891
2 0.500000 0.134871582326500
-N  0
N  3 0
l
7
2 0.120000 -0.046687263525527
2 0.140000 0.112197597324783
2 0.200000 -0.242349891313031
2 0.260000 0.287679595000293
2 0.500000 -0.228734330552699
2 0.900000 0.204445025863325
2 3.000000 0.104721265584314
s
3
2 0.120000 0.059230506441412
2 0.240000 -0.199637443278969
2 1.500000 0.360974224566524
p
3
2 0.120000 -0.050216559697878
2 0.260000 0.120391744221277
2 0.700000 -0.120795692498127
d
3
2 0.120000 0.104840772118806
2 0.200000 -0.235417307068532
2 0.500000 0.152022842490471
-O  0
O  3 0
l
3
2 0.120000 0.001588481299492
2 0.220000 -0.014544876778442
2 0.300000 0.046533571529401
s
3
2 0.120000 0.030966247566569
2 0.220000 -0.005930784430083
2 1.200000 0.011108361287909
```

```
p
4
2 0.140000 0.013931039867617
2 0.240000 -0.033241914562671
2 0.500000 -0.096441224606502
2 1.500000 0.281283766525674
d
3
2 0.140000 0.008525208858068
2 0.280000 -0.140769076716143
2 0.700000 0.103496603229196
-F  0
F  3 0
l
6
2 0.120000 -0.001592887525476
2 0.180000 -0.006849310146700
2 0.280000 0.053111927773545
2 0.300000 0.023035953581296
2 0.500000 -0.144041125832224
2 1.700000 0.073763307215227
s
3
2 0.120000 0.108419697594934
2 0.140000 0.056885614671750
2 0.260000 -0.288522399976770
p
5
2 0.120000 0.027387308748060
2 0.220000 -0.096573363482719
2 0.260000 -0.009673427823668
2 1.200000 0.130295543506617
2 3.000000 0.263207758684173
d
2
2 0.200000 -0.034036999589265
2 0.800000 0.145900599206114
-Si 0
Si 3 0
l
2
2 0.140000 -0.003451594717329
2 0.240000 0.006490192150505
s
1
2 0.180000 -0.005696183993748
p
2
2 0.120000 0.020714788587574
2 0.220000 -0.044664716509119
d
2
2 0.120000 0.021419158226661
2 0.220000 -0.084614839996599
-P  0
P  3 0
l
4
2 0.120000 -0.009601899889960
2 0.160000 0.011591302357969
2 0.180000 0.016142896394789
2 0.400000 -0.092691232840389
s
1
```

```
2 0.120000 0.061880629394294
p
2
2 0.120000 0.022714848707400
2 0.220000 -0.014088421938289
d
3
2 0.160000 -0.003485710051989
2 0.180000 -0.042555256276778
2 0.240000 -0.000000085952109
-S  0
S  3 0
l
5
2 0.120000 -0.004513630398870
2 0.140000 -0.004949059638582
2 0.220000 0.065598336534456
2 0.300000 -0.064116529081196
2 0.600000 -0.051325888233958
s
3
2 0.160000 0.069398465475923
2 0.240000 -0.000007180218602
2 0.300000 0.037318766238815
p
4
2 0.120000 0.081876467199086
2 0.160000 -0.069674222495886
2 0.220000 -0.024229389326715
2 0.700000 0.026011806031683
d
3
2 0.140000 -0.032301038485482
2 0.240000 -0.079416115336865
2 1.100000 -0.000000873697740
-Cl 0
Cl 3 0
l
6
2 0.120000 -0.013272064728508
2 0.140000 0.029454689743362
2 0.180000 -0.010769532165205
2 0.400000 0.081774805547980
2 0.900000 -0.148604283716178
2 1.300000 -0.243898398932391
s
1
2 0.120000 0.119917901605640
p
2
2 0.120000 0.016331479398588
2 0.200000 -0.015799136049257
d
2
2 0.180000 -0.115011165078866
2 0.200000 -0.098333805887596
```

Section S5. ACP file for CAM-B3LYP-D3/6-31G*

```
-H  0
H  1 0
l
9
```

```
2 0.120000 0.003048224284080
2 0.140000 0.015515553678735
2 0.160000 -0.045934470215758
2 0.240000 0.071223103662569
2 0.400000 -0.153754647358492
2 0.600000 0.268842592121685
2 1.000000 -0.312871516974197
2 1.100000 -0.039758332386013
2 2.000000 0.215948557667946
s
5
2 0.140000 0.047879076113302
2 0.200000 0.004819627813731
2 0.600000 -0.272053204537317
2 1.200000 0.136322152203351
2 1.900000 0.281540825816131
-B  0
B  3 0
l
3
2 0.120000 0.001184361997233
2 0.200000 -0.000403690370693
2 0.400000 0.001546838814357
s
1
2 0.220000 -0.043752098629545
p
1
2 0.120000 0.003209859966963
d
2
2 0.120000 -0.010961464497040
2 0.140000 -0.004139540217291
-C  0
C  3 0
l
8
2 0.120000 0.004784238201152
2 0.140000 -0.020130928164277
2 0.180000 0.030830829138640
2 0.260000 -0.006918576304240
2 0.400000 -0.040779987131676
2 0.700000 0.066348630530823
2 1.500000 -0.219646982288927
2 2.500000 -0.148289252827141
s
3
2 0.200000 0.005023321110707
2 0.260000 0.321912258355584
2 0.700000 -0.521302386254849
p
4
2 0.120000 -0.040001472103301
2 0.260000 0.022447601883622
2 0.800000 0.148362181586670
2 1.300000 0.096452558512378
d
3
2 0.120000 0.048578019941934
2 0.160000 -0.087068093152008
2 0.260000 0.029823068892036
-N  0
N  3 0
l
```

```
6
2 0.120000 -0.012455576165504
2 0.140000 0.018640853175512
2 0.180000 -0.015904022755894
2 0.300000 0.042617283086109
2 0.700000 -0.082357169738921
2 2.500000 0.340952914966344
s
2
2 0.200000 -0.036066094826990
2 0.800000 0.052327850481324
p
2
2 0.120000 0.014711853304275
2 0.180000 -0.015601233607044
d
3
2 0.120000 0.096878803529400
2 0.160000 -0.027704060973933
2 0.180000 -0.142384859275971
-O  0
O  3 0
l
5
2 0.120000 -0.002695664420530
2 0.140000 -0.000189367684468
2 0.180000 0.006346714119468
2 0.300000 0.015740573919481
2 0.700000 0.078693501625984
s
3
2 0.120000 0.127909726649770
2 0.260000 -0.317607954097510
2 1.600000 0.340200286447476
p
4
2 0.120000 0.048715904793784
2 0.220000 -0.069153727227757
2 0.600000 -0.139050439045852
2 1.900000 0.323552667083446
d
3
2 0.120000 0.030172850958866
2 0.160000 -0.100521112805213
2 0.700000 -0.090579528756053
-F  0
F  3 0
l
7
2 0.140000 -0.003387420548115
2 0.160000 -0.009227553631959
2 0.260000 0.064998918034252
2 0.400000 -0.028819273303491
2 0.600000 -0.088302796751152
2 1.600000 0.038418383641362
2 3.000000 0.531774770365069
s
3
2 0.120000 0.124674057735850
2 0.300000 -0.114181513386130
2 0.800000 -0.186753202225978
p
2
2 0.120000 0.032579668544353
```

```
2 0.280000 -0.098831500551372
d
3
2 0.120000 0.041451183044098
2 0.200000 -0.161769249010498
2 0.600000 0.225223332611439
-Si 0
Si 3 0
l
3
2 0.120000 -0.001398960773084
2 0.140000 -0.001205261574101
2 0.300000 -0.009409181250650
s
1
2 0.120000 0.002571467237452
p
2
2 0.120000 0.046755484073788
2 0.240000 -0.069328710946667
d
1
2 0.220000 -0.032799511998850
-P  0
P  3 0
l
4
2 0.120000 -0.000632217521208
2 0.140000 -0.008332446278837
2 0.200000 0.029483744737607
2 0.400000 -0.097035105513294
s
1
2 0.120000 0.029027756302283
p
2
2 0.120000 0.001465257087695
2 0.600000 0.080762447586952
d
2
2 0.120000 0.011630737462854
2 0.220000 -0.072712106280418
-S  0
S  3 0
l
3
2 0.120000 -0.004234996303686
2 0.180000 0.002098448955270
2 0.280000 0.009621498100372
s
1
2 0.140000 0.043730377998570
p
2
2 0.120000 0.004042001286084
2 1.200000 -0.004153820984203
d
3
2 0.160000 0.060884370380283
2 0.300000 -0.201682411063981
2 0.600000 -0.296000026616047
-Cl 0
Cl 3 0
l
```

```
5
2 0.120000 -0.001962424802554
2 0.140000 -0.005235219791587
2 0.220000 0.025163072134977
2 0.240000 0.018909415304336
2 1.600000 -0.232538734801448
s
3
2 0.120000 0.228263075761020
2 0.700000 -0.026550632398198
2 0.800000 -0.036384965221715
p
1
2 0.120000 0.002583361871362
d
3
2 0.120000 0.077111635705567
2 0.160000 -0.287088730469376
2 0.500000 0.015381810979171
```

Section S6. Formulas for all the statistical error measures

o) Mean absolute error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where, $x_i = \left|x_{calc,i} - x_{ref,i}\right|$

p) Mean signed error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where, $x_i = x_{calc,i} - x_{ref,i}$

q) Maximum absolute error (MAXE)

$$MAXE = \max_{i} \left|x_{calc,i} - x_{ref,i}\right|$$

r) Root-mean-square error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

where, $x_i = x_{calc,i} - x_{ref,i}$

s) Standard deviation (SD)

$$SD = \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

where, $x_i = x_{calc,i} - x_{ref,i}$

$$\overline{x} = \frac{1}{n}(x_{calc,i} - x_{ref,i})$$

Table S1. Detailed list of data sets in the ACP training set.

Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
		Non-covalent interaction energies:		
HBC6	118	Interaction energies of doubly hydrogen-bonded dimer complexes at various intermolecular distances	CCSD(T)/CBS	77, 78
MiriyalaHB104	104	Interaction energies of hydrogen-bonded dimer complexes	CCSD(T)/CBS	79, 80
IonicHB	96	Interaction energies of charged (both positive and negative) hydrogen-bonded dimer complexes at various intermolecular distances	CCSD(T)/CBS	81
HB375x10	3749	Interaction energies of neutral hydrogen-bonded dimer complexes at various intermolecular distances	CCSD(T)/CBS	82
IHB100x10	350	Interaction energies of charged (both positive and negative) hydrogen-bonded dimer complexes at various intermolecular distances	CCSD(T)/CBS	82
HB300SPXx10	1980	Interaction energies of neutral hydrogen-bonded dimer complexes in an extended chemical space (excluding iodine and bromine containing systems) at various intermolecular distances.	CCSD(T)/CBS	83
CARBHB12	12	Interaction energies of dimer complexes between singlet carbene analogues and H_2O, NH_3, HCl	W2-F12	14
S22x5	110	Interaction energies of small non-covalently bound dimer complexes at various intermolecular distances	CCSD(T)/CBS	78,84,85
S66x8	528	Interaction energies of small non-covalently bound dimer complexes at various intermolecular distances	CCSD(T)/CBS	86–88
S66a8	528	Interaction energies of small non-covalently bound dimer complexes at various intermolecular angular displacements	CCSD(T)/CBS	87
A21x12	228	Interaction energies of small non-covalently bound dimer complexes (excluding argon containing systems) at various intermolecular distances	CCSD(T)/CBS	89–91
NBC10ext	195	Interaction energies of non-covalently interacting dimer complexes at various intermolecular distances	CCSD(T)/CBS	78,92–95
3B-69-DIM	207	Interaction energies of all relevant pairs of monomers from 3B-69-TRIM	CCSD(T)/CBS	96
3B-69-TRIM	69	Interaction energies of trimer complexes of small organic molecules	CCSD(T)/CBS	96

Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
HW30	30	Interaction energies of dimer complexes of hydrocarbons and water	CCSD(T)/CBS	97
B-set	160	Interaction energies of dimer complexes containing boron at various intermolecular distances	DLPNO-CCSD(T)/CBS	58, 64
F-set	160	Interaction energies of dimer complexes containing fluorine at various intermolecular distances	DLPNO-CCSD(T)/CBS	58, 64
Si-set	152	Interaction energies of dimer complexes containing silicon at various intermolecular distances	DLPNO-CCSD(T)/CBS	58, 64
P-set	120	Interaction energies of dimer complexes containing phosphorus at various intermolecular distances	DLPNO-CCSD(T)/CBS	58, 64
S-set	144	Interaction energies of dimer complexes containing sulfur at various intermolecular distances	DLPNO-CCSD(T)/CBS	58, 64
Cl-set	160	Interaction energies of dimer complexes containing chlorine at various intermolecular distances	DLPNO-CCSD(T)/CBS	58, 64
Sulfurx8	104	Interaction energies of dimer complexes containing divalent sulfur at various intermolecular distances	CCSD(T)/CBS	98
Pisub	105	Interaction energies of non-covalently bound substituted aromatic dimer complexes	DLPNO-CCSD(T)/CBS	58, 99, 100
Pi29n	29	Interaction energies of neutral π-conjugated dimer complexes representing organic electronic precursors	CCSD(T)/CBS	101
BzDC215	170	Interaction energies of dimer complexes (excluding neon and argon containing systems) of benzene with small molecules	CCSD(T)/CBS	102
C2H4NT	75	Interaction energies of dimer complexes of ethene and coronene	CCSD(T)/CBS	95
Hill18	18	Interaction energies of hydrogen-bonded and halogen-bonded dimer complexes	CCSD(T)/CBS	103
X40x10	220	Interaction energies of dimer complexes (excluding iodine and bromine containing systems) representing halogen-bonding at various intermolecular distances	CCSD(T)/CBS	104
PNICO23	23	Interaction energies of dimer complexes representing pnictogen-bonding	W1-F12, W2-F12	14, 105
ADIM6	6	Interaction energies of six alkane dimer complexes ranging from ethane to n-heptane	W1-F12	14, 25, 106
HC12	12	Interaction energies of saturated and unsaturated hydrocarbon dimer complexes	CCSD(T)/CBS	107
BBI	100	Interaction energies of peptide backbone-backbone dimer complexes	DW-CCSD(T)-F12/aug-cc-pV(D+d)z	108
SSI	2805	Interaction energies of peptide sidechain-sidechain dimer complexes	DW-CCSD(T)-F12/aug-cc-pV(D+d)z	108

Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
NucTAA	454	Interaction energies of dimer complexes of amino acid and nucleotide	DLPNO-CCSD(T)/CBS	58, 109–112
CarbhydBz	34	Interaction energies of carbohydrate-benzene dimer complexes	CCSD(T)/CBS	113
CarbhydNaph	46	Interaction energies of carbohydrate-naphthalene dimer complexes	CCSD(T)/CBS	114
CarbhydAroAA	48	Interaction energies of dimer complexes representing carbohydrate and aromatic amino acids	DLPNO-CCSD(T)/CBS	58, 115
CarbhydAro	161	Interaction energies of dimer complexes representing carbohydrate and substituted aromatic molecule	DLPNO-CCSD(T)/CBS	58, 116
WatAA	259	Interaction energies of dimer complexes representing interactions between water and amino acids	DLPNO-CCSD(T)/CBS	58, 117
HSG	17	Interaction energies of dimer complexes representing protein-ligand interactions	CCSD(T)/CBS	78, 118
PLF547	392	Interaction energies of dimer complexes representing protein-ligand interactions	DLPNO-CCSD(T)/CBS	119
JSCH	124	Interaction energies of dimer complexes of nucleotide base pairs	CCSD(T)/CBS	84
DNAstack	40	Interaction energies of stacked DNA base-pair steps	CBS(T)-F12-CP	120
DNA2body	10	Interaction energies of nucleobase pairs	CBS(T)-F12-CP	120
ACHC	54	Interaction energies of nucleobase stacking configurations	DW-CCSD(T**)-F12/aug-cc-pVDZ	121
BDNA	71	Interaction energies of nucleobase stacking configurations	CCSD(T)/CBS	122
NucBTrimer	141	Interaction energies of complexes of nucleobase trimers	DLPNO-CCSD(T)/CBS	58, 123
Water38	38	Interaction energies of water clusters $(H_2O)_n$ (where n = 2-10)	CCSD(T)/CBS	124
Water1888	1888	Interaction energies of various water dimer configurations with reference data lying between -5 to +5 kcal/mol	CCSD(T)/CBS	95,125–127
Water-2body	410	Interaction energies of various water dimer configurations	CCSD(T)/CBS	54, 58
CH4PAH	382	Interaction energies of dimer complexes of methane and polycyclic aromatic hydrocarbons	CCSD(T)/CBS	128, 129
CO2MOF	20	Interaction energies of dimer complexes of carbon dioxide and organic building units of metal-organic frameworks	inc-CCSD(T)|MP2+F12+INT/cc-pVDZ-F12	130
CO2PAH	249	Interaction energies of dimer complexes of carbon dioxide and polycyclic aromatic hydrocarbons	CCSD(T**)-F12avg/CBS	131
CO2NPHAC	96	Interaction energies of dimer complexes of carbon dioxide and nitrogen-doped poly-heterocyclic aromatic compounds	CCSD(T)/CBS	132
BzGas	129	Interaction energies of nine benzene-gas dimer complexes at various intermolecular distances (where gas = CO_2, CH_4, N_2)	CCSD(T)/CBS	133

Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
SSI-anionic, WatAA-anionic, HSG-anionic, PLF547-anionic, IonicHB-anionic, IHB100x10-anionic	575, 64, 4, 155, 24, 650	Interaction energies of only anion-neutral and anion-cation dimer complexes from earlier described datasets	Various	58, 78, 81, 82, 108, 117–119
Ionic43-anionic	37	Interaction energies of anion-neutral and anion-cation dimer complexes (excluding sodium, potassium, and lithium containing complexes)	CCSD(T)/CBS	134
Molecular conformational energies:				
37Conf8	258	Relative energies of conformers of organic molecule isomers	DLPNO-CCSD(T)/cc-pVTZ	135
DCONF	2142	Relative energies of conformers of 62 model systems representing drug-like molecules at various intramolecular torsion angles	CCSD(T)/CBS	136
ICONF	17	Relative energies of conformers of 10 molecules containing H, N, O, Si, P, and S	W1-F12	14
MCONF	51	Relative energies of conformers of melatonin	CCSD(T)/CBS	137
Torsion21	189	Relative energies of conformers of Glyoxal, Oxalyl halides, and their thiocarbonyl derivatives (excluding bromine containing systems) at various intramolecular torsion angles	CCSD(T)/CBS	138
MolCONF	5623	Relative energies of conformers of molecules taken from crystal structure database and protein-ligand database (only containing our 10 target elements)	DLPNO-CCSD(T)/cc-pVTZ	139
PEPCONF-Dipeptide	875	Relative energies of conformers of various model dipeptide systems	DLPNO-CCSD(T)/CBS	58, 140
TPCONF	8	Relative energies of conformers of two model tetrapeptides	CCSD(T)/CBS	141
P76	71	Relative energies of conformers of five isolated small peptides containing aromatic side chains	CCSD(T)/CBS	142
YMPJ	495	Relative energies of conformers of proteinogenic amino acid monomers	MP2-F12/cc-pVTZ-F12+[CCSD(Ts)-F12b – MP2-F12]/cc-pVDZ-F12	143
SPS	17	Relative energies of conformers of DNA sugar-phosphate-sugar backbone	CCSD(T)/CBS	144
rSPS	45	Relative energies of conformers of RNA sugar-phosphate-sugar backbone	CCSD(T)/CBS	145
UpU46	45	Relative energies of conformers of model RNA backbone	DLPNO-CCSD(T)/CBS*	146
SCONF	17	Relative energies of conformers of two model carbohydrates	CCSD(T)/CBS	14, 147
DSCONF	27	Relative energies of conformers of three disaccharides	CCSD(T)/CBS	148
SacchCONF	56	Relative energies of conformers of monosaccharides	CCSD(T)/CBS	149
CCONF	426	Relative energies of conformers of glucose and α-maltose isomers	DLPNO-CCSD(T)/CBS	150

Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
ACONF	15	Relative energies of conformers of n-alkane chains	W1h-val	151
BCONF	64	Relative energies of conformers of butane-1,4-diol	CCSD(T)-F12b/cc-pVTZ-F12	152
PentCONF	342	Relative energies of conformers of n-pentane	CCSD(T)-F12/CBS	153
Undecamer125	124	Relative energies of conformers of $(H_2O)_{11}$	CCSD(T)/CBS	154
PEPCONF-Dipeptide-anionic, MolCONF-anionic	175, 79	Relative energies of conformers of systems containing only negative charge from earlier described datasets	Various	58, 139, 140
Reaction energies:				
MN-RE	7555	Automatically generated reactions using molecules from Minnesota Database2015B[156]	Various	155
BH9-RE	449	From BH9 set comprising chemical reactions belonging to nine types common in organic chemistry and biochemistry	DLPNO-CCSD(T)/CBS	74
DIE60	60	Double-bond migration reactions in conjugated dienes	W*n*-F12	157
FH51	51	Reactions involving various organic and inorganic molecules	CCSD(T)-F12/CBS	158, 159
BSR36	36	Hydrocarbon bond separation reactions	CCSD(T)/CBS	160, 161
BH76RC	30	Hydrogen and non-hydrogen atom transfer reactions of small molecules	W*n*	162–164
G2RC	23	Reactions whose reactants and products had been taken from the G2/97 set	W2-F12	141, 164, 165
RC21	21	Organic radical fragmentation and rearrangement reactions	W1-F12	14
CR20	20	Cyclo-reversion reactions	W1-F12	166
PlatonicHD6	6	Homodesmotic reactions involving platonic hydrocarbon cages, C_nH_n (where n = 4,6,8,10,12,20)	W*n*-F12	167
PlatonicID6	6	Isodesmic reactions involving platonic hydrocarbon cages, C_nH_n (where n = 4,6,8,10,12,20)	W*n*-F12	167
PlatonicIG6	6	Isogyric reactions involving platonic hydrocarbon cages, C_nH_n (where n = 4,6,8,10,12,20)	W*n*-F12	167
AlkIsod14	14	Isodesmic reactions involving C_nH_{2n+2} alkanes (where n=3-8)	W*n*	168
DARC	14	Diels-Alder reactions	W1-F12	14, 164, 169
DC13	12	Reactions that were known to be difficult for DFT methods	Various	14,63,178,179,170–177
WCPT6	6	Tautomeric water-catalyzed proton transfer reactions	W*n*	180
NBPRC	6	Reactions involving NH_3/BH_3 and PH_3/BH_3	CCSD(T)/CBS	161, 164, 181
Barrier height energies:				

Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
Grambow2020-B97D3	32722	Reactions involving H, C, N, and O generated using automated potential energy surface exploration	DLPNO-CCSD(T)/CBS	182, This work
Grambow2020-ωB97XD3	23922	Reactions involving H, C, N, and O generated using automated potential energy surface exploration	DLPNO-CCSD(T)/CBS	182, This work
BH9	898	Chemical reactions belonging to nine types common in organic chemistry and biochemistry	DLPNO-CCSD(T)/CBS	74
E2SN2	418	Competing E2 and S_N2 reactions	DLPNO-CCSD(T)/CBS	183, This work
HTBH38	38	Hydrogen atom transfer reactions of small molecules	W1 and theoretical estimate	163
NHTBH38	38	Non-hydrogen atom transfer reactions of small molecules	W1 and theoretical estimate	162
WCPT27	27	Water-catalyzed proton-transfer reactions	W*n*	180
BHROT27	27	Rotation around single bonds	W*n*-F12	14
BHPERI26	26	Pericyclic reactions	W*n*-F12	164, 184
DBH24	24	Diverse reactions involving small molecules	W1 and theoretical estimate	185, 186
INV24	24	Inversion and racemization reactions	W*n*-F12 and DLPNO-CCSD(T)/CBS	187
CRBH20	20	Cyclo-reversion reactions of heterocyclic rings	W*n*	188
PX13	13	Proton exchange reactions in small clusters of H_2O, NH_3, and HF	W1-F12	189, 190
Bond separation energies:				
BSE49	4502	Breaking of 49 unique X-Y type single bonds (except H-H, H-F, and H-Cl) into corresponding radical fragments, where X and Y are H, B, C, N, O, F, Si, P, S, Cl	(RO)CBS-QB3	76
Molecular deformation energies:				
MOLdef	9298	Molecular deformation energies relative to the equilibrium geometry of systems containing our 10 target elements	DLPNO-CCSD(T)/CBS	43, 58
MOLdef-H2O	990	Molecular deformation energies relative to the equilibrium geometry of water containing systems	CCSD(T)/CBS	58, 191, 192
MOLdef-TS	6294	Molecular deformation energies relative to the transition structure of BH9 data set. Molecules were deformed along the imaginary normal mode only.	DLPNO-CCSD(T)/CBS	74, This work
Isomerization energies:				
ISO34	34	Relative energies of isomers of small and medium-sized organic molecules	W1-F12	14, 193
ISOL24	24	Relative energies of isomers of large organic molecules	DLPNO-CCSD(T)/CBS	14, 194
IDISP	6	Relative energies of isomers of hydrocarbon molecules	DLPNO-CCSD(T)/CBS	14, 161, 164, 193, 195, 196
EIE22	22	Relative energies of isomers of enecarbonyls	W1-F12	197
PArel	20	Relative energies of protonated isomers	CCSD(T)/CBS	14

Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
AlkIsomer11	11	Relative energies of isomers of C_nH_{2n+2} where n=4–8	W*n*	168
PAH6	6	Relative energies of polycyclic aromatic hydrocarbon isomers	CCSD(T)/CBS	198
Styrene45	44	Relative energies of isomers of C_8H_8	W1-F12	170
TAUT15	15	Relative energies in tautomeric molecules	W1-F12	14
H2O16Rel15	4	Relative energies of isomers of $(H_2O)_{16}$ (boat and fused cube structures)	CCSD(T)/aug-cc-pVTZ	199
H2O20Rel10	9	Relative energies of isomers of $(H_2O)_{20}$ (lowest-energy structures)	CCSD(T)/CBS	200
SW49Rel6	17	Relative energies of isomers of $SO_4^{2-}(H_2O)_6$	CCSD(T)/CBS	201
SW49Rel345	28	Relative energies of isomers of $SO_4^{2-}(H_2O)_n$ where n=3–5	CCSD(T)/CBS	201
Total atomization energies:				
W4-17	194	First- and second-row molecules and radicals with up to eight non-hydrogen atoms	W4	202
PlatonicTAE6	6	Platonic hydrocarbon cages, C_nH_n (where n = 4,6,8,10,12,20)	W*n*-F12	167
AlkAtom19	19	C_nH_{2n+2} where n=1–8	W*n*	168

Table S2. Detailed list of data sets in the ACP validation set.

Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
Non-covalent interaction energies:				
BlindNCI	80	Interaction energies of 10 dimer complexes at various intermolecular distances used previously for blind test	CCSD(T)/CBS	203
DES15K	11474	Interaction energies of various non-covalently bound dimer complexes	CCSD(T)/CBS	204
NENCI-2021	5859	Interaction energies of non-equilibrium dimer complexes	CCSD(T)/CBS	205
CE20	20	Interaction energies of water, ammonia, and hydrogen fluoride clusters	W1-F12	189, 190
WaterOrg	2,376	Interaction energies of hydrogen-bonding interactions between water clusters and organic molecule complexes	DLPNO-CCSD(T)/CBS	206
R160x6	960	Interaction energies of small dimer complexes at short intermolecular distances	CCSD(T)/CBS	207
R739x5	4330	Interaction energies of small dimer complexes at short intermolecular distances	CCSD(T)/CBS	208
CHAL336	48	Interaction energies of chalcogen-bonded dimer complexes	W1-F12 or DLPNO-CCSD(T)/CBS	209
XB45	33	Interaction energies of halogen-bonded dimer complexes	CCSD(T)/aug-cc-pVTZ	210
L7	7	Interaction energies of seven relatively large non-covalently bound complexes	DLPNO-CCSD(T)/CBS	211, 212

Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
S12L	10	Interaction energies of supramolecular host-guest complexes	DLPNO-CCSD(T)/CBS	9, 11, 212
S30L	26	Interaction energies of supramolecular host-guest complexes	Experimental back-corrected	213
Ni2021	11	Interaction energies of large non-covalently bound complexes ranging in size between 126-1027 atoms	CIM-DLPNO-CCSD(T)||RI-MP2	214
C60dimer	14	Interaction energies of the C_{60} dimer complex at various intermolecular distances	DLPNO-CCSD(T)/CBS	215
H2O20Bind10	10	Interaction energies of clusters of $(H_2O)_{20}$	CCSD(T)/CBS	200
HW6Cl-anionic	6	Interaction energies of clusters of $Cl^-(H_2O)_n$ (where n = 1-6)	CCSD(T)/CBS	200, 216
HW6F-anionic	6	Interaction energies of clusters of $F^-(H_2O)_n$ (where n = 1-6)	CCSD(T)/CBS	200, 216
FmH2O10-anionic	10	Interaction energies of clusters of $F^-(H_2O)_{10}$	CCSD(T)/CBS	200, 216
SW49Bind345-anionic	30	Interaction energies of clusters of $SO_4^{2-}(H_2O)_n$ (where n = 3-5)	CCSD(T)/CBS	201
SW49Bind6-anionic	18	Interaction energies of clusters of $SO_4^{2-}(H_2O)_6$	CCSD(T)/CBS	201
Anionpi-anionic	16	Interaction energies of anion-π type non-covalently interacting dimer complexes	DLPNO-CCSD(T)/CBS	217
IL236-anionic	236	Interaction energies of ion pair dimer complexes representing model ionic liquids	CCSD(T)/CBS	218
DES15K-anionic, NENCI-2021-anionic, CHAL336-anionic, XB45-anionic, S30L-anionic	1281, 889, 19, 12, 2	Interaction energies of only anion-neutral and anion-cation dimer complexes from earlier described datasets	Various	204, 205, 209, 210, 213
Molecular conformational energies:				
SafroleCONF	5	Relative energies of safrole conformers	CCSD(T)/CBS	219
AlcoholCONF	31	Relative energies of small alcohol conformers	CCSD(T)/aug-cc-pVTZ	220
BeranCONF	50	Relative energies of flexible organic molecule conformers relevant in crystal structure prediction	CCSD(T)/CBS or MP2D/CBS	221
Torsion30	2107	Relative energies of conformers of model systems representing biaryl drug-like molecules at various intramolecular torsion angles (excluding many datapoints for which geometries were missing).	CCSD(T)*/CBS	222
ANI1ccxCONF	5254	Relative energies with respect to the energy minimum of organic molecules generated using normal mode sampling, dimer sampling, and torsion sampling	CCSD(T)*/CBS	223
MPCONF196	112	Relative energies of medium-sized macrocyclic peptide conformers	DLPNO-CCSD(T)/CBS	224
PEPCONF-Tripeptide	647	Relative energies of conformers of various model tripeptide systems	DLPNO-CCSD(T)/CBS	58, 140
PEPCONF-Disulfide	620	Relative energies of conformers of various model peptide systems containing disulfide linkages	LC-ωPBE-XDM/aug-cc-pVTZ	140
PEPCONF-Cyclic	320	Relative energies of conformers of various macrocyclic peptides	LC-ωPBE-XDM/aug-cc-pVTZ	140
PEPCONF-Bioactive	175	Relative energies of conformers of various polypeptides that show bioactive function	LC-ωPBE-XDM/aug-cc-pVTZ	140

Data set(s)	Data points	Data set description	Reference energy level	Reference # in article
PEPCONF-Disulfide-anionic, PEPCONF-Bioactive-anionic	150, 20	Relative energies of conformers of systems containing only negative charge from earlier described datasets	LC-ωPBE-XDM/aug-cc-pVTZ	140
Reaction energies:				
W4-17-RE	5205	Automatically generated reactions using molecules from the W4-17[202]	W4	155
Barrier height energies:				
WaterOrgBH	88	Pericyclic reactions in absence and presence of water clusters	DLPNO-CCSD(T)/CBS	unpublished data

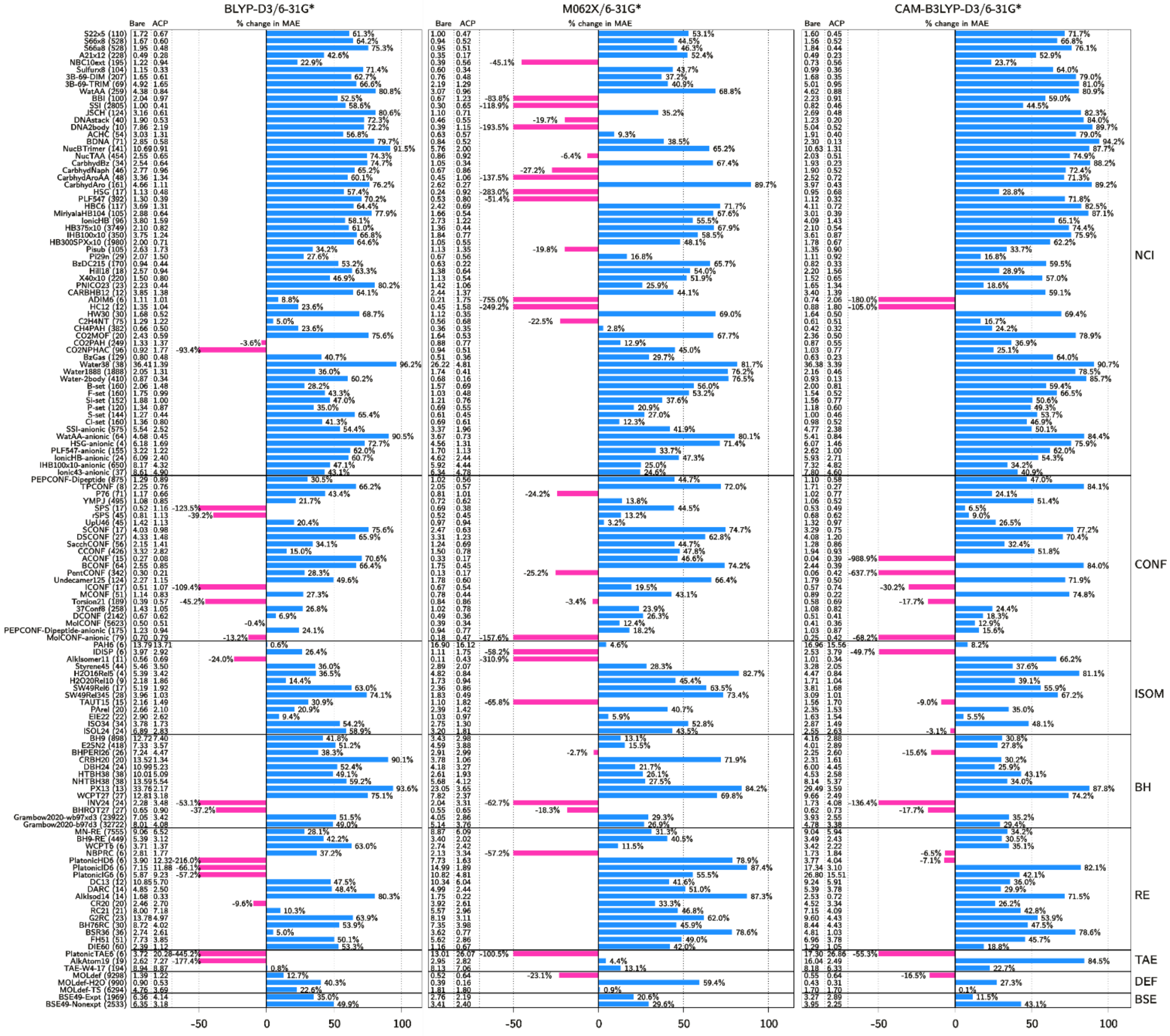


Figure S1. Mean absolute errors (MAEs, in kcal/mol) of uncorrected methods (“Bare”) and ACP-corrected methods (“ACP”) along with percentage change in MAEs on application of ACPs to the training set. The various shorthand notations are as follows: NCI = non-covalent interaction energies, CONF = molecular

conformational energies, ISOM = isomerization energies, BH = barrier heights, RE = reaction energies, TAE = total atomization energies, DEF = molecular deformation energies, and BSE = bond separation energies. The % change in MAE is calculated as [MAE of ACP-corrected method] – [MAE of uncorrected method] / [MAE of uncorrected method] x 100%.

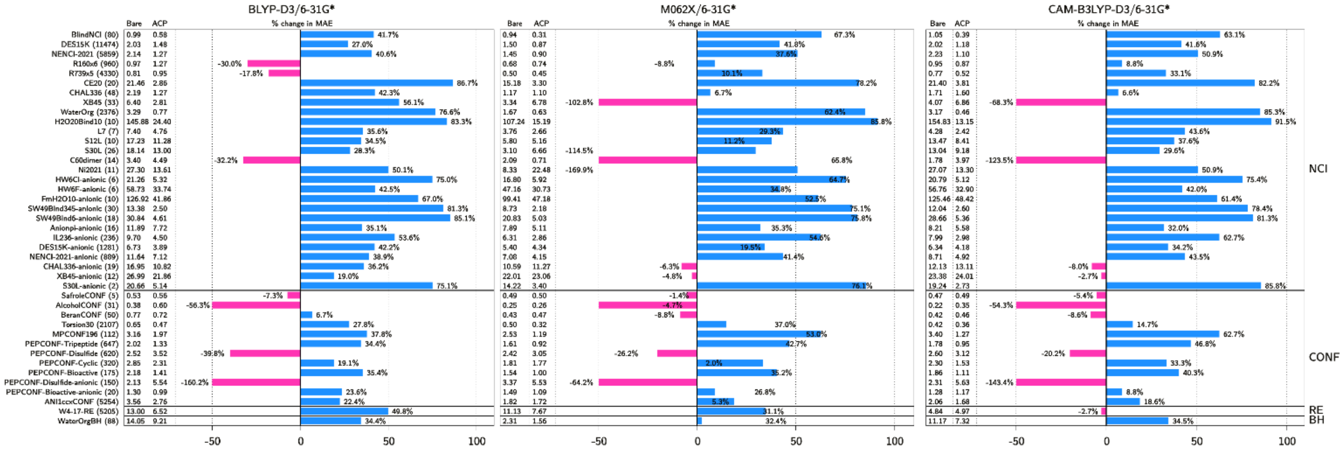


Figure S2. Mean absolute errors (MAEs) of uncorrected methods (“bare”) and ACP-corrected methods (“ACP”) along with percentage change in MAEs on application of ACPs to the validation set. The various shorthand notations are as follows: NCI = non-covalent interaction energies, CONF = molecular conformational energies, RE = reaction energies, and BH = barrier heights. The % change in MAE is calculated as [MAE of ACP-corrected method] – [MAE of uncorrected method] / [MAE of uncorrected method] x 100%.

Table S3. Detailed error analysis with respect to reference data in the training set. The numbers in bracket in the first column indicates the number of data points. The various shorthand notations are as follows: MAE = mean absolute error in kcal/mol, MSE = mean signed error in kcal/mol, MAXE = maximum absolute error in kcal/mol, RMSE = root-mean-square error in kcal/mol, and SD = standard deviation in kcal/mol.

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
S22x5 (110)	MAE	1.72	0.67	1.00	0.47	1.60	0.45
	MSE	-1.72	-0.38	-0.75	0.12	-1.59	-0.20
	MAXE	5.73	5.08	4.05	2.49	6.87	2.30
	RMSE	2.46	1.09	1.45	0.66	2.48	0.68
	SD	1.77	1.03	1.25	0.65	1.91	0.65
S66x8 (528)	MAE	1.67	0.60	0.94	0.52	1.56	0.52
	MSE	-1.67	-0.12	-0.79	0.22	-1.56	0.03
	MAXE	5.57	5.33	3.48	2.82	6.47	4.25
	RMSE	2.08	0.90	1.25	0.77	2.08	0.77
	SD	1.23	0.89	0.98	0.74	1.38	0.77
S66a8 (528)	MAE	1.95	0.48	0.95	0.51	1.84	0.44
	MSE	-1.95	-0.10	-0.81	0.32	-1.84	-0.04
	MAXE	5.62	2.35	3.83	2.10	6.43	1.77
	RMSE	2.18	0.61	1.28	0.67	2.18	0.55
	SD	0.96	0.60	0.99	0.59	1.16	0.55
A21x12 (228)	MAE	0.49	0.28	0.35	0.17	0.49	0.23

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
	MSE	-0.47	-0.07	-0.32	-0.04	-0.48	-0.11
	MAXE	4.72	3.09	3.62	1.71	5.54	2.24
	RMSE	0.96	0.61	0.72	0.29	1.01	0.43
	SD	0.84	0.61	0.65	0.29	0.88	0.42
NBC10ext (195)	MAE	1.22	0.94	0.39	0.56	0.73	0.56
	MSE	-1.22	-0.89	-0.11	-0.26	-0.73	-0.45
	MAXE	3.25	3.74	1.44	2.26	1.40	1.37
	RMSE	1.56	1.31	0.52	0.80	0.88	0.70
	SD	0.97	0.97	0.51	0.76	0.49	0.54
Sulfurx8 (104)	MAE	1.15	0.33	0.60	0.34	0.99	0.36
	MSE	-1.15	-0.08	-0.54	0.10	-0.99	-0.07
	MAXE	3.07	1.53	2.15	1.20	3.00	1.42
	RMSE	1.43	0.44	0.82	0.44	1.28	0.49
	SD	0.86	0.44	0.61	0.43	0.80	0.48
3B-69-DIM (207)	MAE	1.65	0.61	0.76	0.48	1.68	0.35
	MSE	-1.64	0.31	-0.69	0.38	-1.67	0.09
	MAXE	6.34	4.16	3.30	2.05	6.44	1.62
	RMSE	2.09	0.99	1.11	0.62	2.18	0.49
	SD	1.29	0.94	0.87	0.48	1.40	0.49
3B-69-TRIM (69)	MAE	4.92	1.65	2.19	1.29	5.01	0.95
	MSE	-4.92	0.92	-2.14	1.08	-5.01	0.27
	MAXE	12.00	6.14	10.01	4.13	12.69	3.59
	RMSE	5.41	2.20	2.76	1.60	5.58	1.21
	SD	2.26	2.01	1.76	1.20	2.47	1.19
WatAA (259)	MAE	4.38	0.84	3.07	0.96	4.62	0.88
	MSE	-4.38	0.08	-3.07	-0.35	-4.62	-0.40
	MAXE	8.70	2.37	7.00	2.51	9.14	2.49
	RMSE	4.57	1.07	3.32	1.13	4.80	1.11
	SD	1.28	1.06	1.26	1.07	1.28	1.04
BBI (100)	MAE	2.04	0.97	0.67	1.23	2.23	0.91
	MSE	-2.04	0.93	-0.64	1.23	-2.23	0.91
	MAXE	2.79	2.89	1.57	2.04	3.19	2.03
	RMSE	2.06	1.13	0.77	1.29	2.31	1.03
	SD	0.30	0.64	0.42	0.38	0.59	0.49
SSI (2805)	MAE	1.00	0.41	0.30	0.65	0.82	0.46
	MSE	-1.00	0.18	-0.12	0.63	-0.82	0.35
	MAXE	5.06	8.23	3.96	3.63	5.29	4.31
	RMSE	1.22	0.59	0.53	0.73	1.08	0.58
	SD	0.71	0.57	0.52	0.36	0.70	0.46
JSCH (124)	MAE	3.16	0.61	1.10	0.71	2.69	0.48
	MSE	-3.16	-0.05	-0.84	0.51	-2.69	0.34
	MAXE	7.75	1.87	4.88	2.41	9.29	1.94
	RMSE	3.68	0.79	1.57	0.90	3.52	0.66
	SD	1.90	0.79	1.33	0.74	2.27	0.56
DNAstack (40)	MAE	1.90	0.53	0.46	0.55	1.23	0.20
	MSE	-1.90	-0.41	0.05	0.35	-1.22	0.13
	MAXE	3.88	1.41	1.14	1.40	2.60	1.12
	RMSE	2.27	0.71	0.52	0.63	1.46	0.33
	SD	1.26	0.59	0.52	0.53	0.82	0.30
DNA2body (10)	MAE	7.86	2.19	0.39	1.15	5.04	0.52
	MSE	-7.86	-2.19	-0.14	1.05	-5.04	0.23
	MAXE	8.80	3.21	0.91	1.90	5.68	1.24
	RMSE	7.87	2.29	0.48	1.28	5.05	0.63

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
	SD	0.42	0.70	0.48	0.77	0.30	0.61
ACHC (54)	MAE	3.03	1.31	0.63	0.57	1.91	0.40
	MSE	-3.03	-1.31	-0.31	-0.38	-1.91	-0.40
	MAXE	4.55	1.76	3.66	1.86	2.26	0.64
	RMSE	3.10	1.34	0.86	0.70	1.94	0.43
	SD	0.62	0.27	0.80	0.59	0.37	0.16
BDNA (71)	MAE	2.85	0.58	0.84	0.52	2.30	0.13
	MSE	-2.85	-0.51	-0.57	0.36	-2.30	0.00
	MAXE	6.06	1.31	2.68	1.05	6.21	0.49
	RMSE	3.45	0.73	1.29	0.55	3.07	0.19
	SD	1.95	0.53	1.16	0.42	2.04	0.19
NucBTrimer (141)	MAE	10.69	0.91	5.76	2.00	10.63	1.31
	MSE	-10.69	0.01	-5.76	1.53	-10.63	1.17
	MAXE	14.95	4.14	9.79	4.58	16.76	4.01
	RMSE	10.84	1.17	6.01	2.30	10.86	1.58
	SD	1.82	1.18	1.72	1.72	2.23	1.07
NucTAA (454)	MAE	2.55	0.65	0.86	0.92	2.03	0.51
	MSE	-2.53	-0.07	-0.45	0.79	-2.02	0.03
	MAXE	13.86	7.50	7.56	3.60	10.01	4.66
	RMSE	3.19	1.00	1.33	1.07	2.61	0.73
	SD	1.94	1.00	1.25	0.73	1.66	0.73
CarbhydBz (34)	MAE	2.54	0.64	1.05	0.34	1.93	0.23
	MSE	-2.54	-0.62	-1.03	0.34	-1.93	0.23
	MAXE	4.30	1.14	2.47	1.32	3.37	1.02
	RMSE	2.62	0.67	1.15	0.44	2.01	0.30
	SD	0.65	0.26	0.53	0.29	0.56	0.21
CarbhydNaph (46)	MAE	2.77	0.96	0.67	0.86	1.90	0.52
	MSE	-2.77	-0.96	-0.62	0.86	-1.90	0.52
	MAXE	3.88	1.53	2.00	1.44	2.72	0.88
	RMSE	2.81	1.00	0.81	0.93	1.95	0.56
	SD	0.51	0.28	0.52	0.37	0.44	0.20
CarbhydAroAA (48)	MAE	3.36	1.34	0.45	1.06	2.52	0.72
	MSE	-3.36	-1.34	0.15	1.06	-2.52	-0.72
	MAXE	6.45	3.27	1.24	1.76	4.88	1.65
	RMSE	3.60	1.49	0.52	1.12	2.70	0.81
	SD	1.30	0.67	0.50	0.36	0.98	0.37
CarbhydAro (161)	MAE	4.66	1.11	2.62	0.27	3.97	0.43
	MSE	-4.66	-1.01	-2.62	0.03	-3.97	-0.36
	MAXE	8.89	2.73	6.32	1.18	7.80	1.43
	RMSE	4.93	1.23	2.93	0.37	4.26	0.51
	SD	1.61	0.70	1.31	0.37	1.56	0.36
HSG (17)	MAE	1.13	0.48	0.24	0.92	0.95	0.68
	MSE	-1.13	0.30	-0.08	0.90	-0.95	0.63
	MAXE	2.65	1.40	1.40	1.90	2.84	2.02
	RMSE	1.27	0.64	0.40	1.02	1.13	0.88
	SD	0.59	0.58	0.40	0.48	0.62	0.64
PLF547 (392)	MAE	1.30	0.39	0.53	0.80	1.12	0.32
	MSE	-1.28	0.14	0.03	0.79	-1.10	0.17
	MAXE	6.45	4.63	3.14	3.41	5.84	1.96
	RMSE	1.83	0.66	0.75	1.01	1.64	0.48
	SD	1.31	0.64	0.75	0.63	1.22	0.45
HBC6 (117)	MAE	3.69	1.31	2.42	0.69	4.11	0.72
	MSE	-3.66	-0.13	-2.39	-0.36	-4.10	-0.59

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
	MAXE	6.92	6.84	5.20	3.61	7.64	2.39
	RMSE	4.37	1.87	2.92	0.95	4.92	0.93
	SD	2.39	1.88	1.69	0.88	2.73	0.72
MiriyalaHB 104(105)	MAE	2.88	0.64	1.66	0.54	3.01	0.39
	MSE	-2.88	0.45	-1.65	0.30	-3.01	0.03
	MAXE	5.53	3.03	3.60	1.74	6.86	1.64
	RMSE	3.01	0.85	1.88	0.66	3.17	0.50
	SD	0.88	0.73	0.91	0.58	1.02	0.50
IonicHB (96)	MAE	3.80	1.59	2.73	1.22	4.09	1.43
	MSE	-3.76	-1.35	-2.72	-1.02	-4.08	-1.36
	MAXE	7.14	4.10	6.44	3.32	7.81	3.23
	RMSE	4.27	1.90	3.14	1.49	4.51	1.68
	SD	2.03	1.34	1.59	1.09	1.93	0.98
HB375x10 (3749)	MAE	2.10	0.82	1.36	0.44	2.10	0.54
	MSE	-2.10	0.35	-1.33	0.16	-2.10	0.06
	MAXE	6.46	9.37	5.64	3.46	7.59	5.08
	RMSE	2.44	1.26	1.68	0.62	2.49	0.73
	SD	1.25	1.21	1.04	0.60	1.35	0.73
IHB100x10 (350)	MAE	3.75	1.24	1.84	0.77	3.61	0.87
	MSE	-3.74	-0.62	-1.81	-0.20	-3.60	-0.65
	MAXE	7.74	6.05	6.52	4.16	7.67	4.32
	RMSE	4.14	1.58	2.24	1.08	3.93	1.14
	SD	1.79	1.45	1.32	1.07	1.56	0.94
HB300SPXx10 (1980)	MAE	2.00	0.71	1.05	0.55	1.78	0.67
	MSE	-1.99	-0.21	-0.98	0.20	-1.77	-0.01
	MAXE	8.54	6.18	6.77	4.33	9.39	4.28
	RMSE	2.48	1.05	1.56	0.80	2.36	0.98
	SD	1.48	1.03	1.22	0.77	1.56	0.98
Pisub (105)	MAE	2.63	1.73	1.13	1.35	1.35	0.90
	MSE	-2.63	-1.72	-0.97	-1.27	-1.35	-0.84
	MAXE	5.49	3.87	2.98	3.14	3.07	2.10
	RMSE	2.83	1.96	1.33	1.63	1.43	1.02
	SD	1.07	0.93	0.92	1.02	0.48	0.58
Pi29n (29)	MAE	2.07	1.50	0.67	0.56	1.11	0.92
	MSE	-2.07	-1.48	0.57	0.02	-1.11	-0.91
	MAXE	7.57	6.55	1.97	3.82	2.07	3.89
	RMSE	2.46	1.96	0.83	0.89	1.19	1.18
	SD	1.34	1.32	0.61	0.90	0.42	0.76
BzDC215 (170)	MAE	0.94	0.44	0.63	0.22	0.82	0.33
	MSE	-0.94	-0.42	-0.59	-0.14	-0.82	-0.27
	MAXE	2.99	1.88	3.17	0.77	3.03	1.33
	RMSE	1.19	0.61	0.93	0.29	1.05	0.43
	SD	0.73	0.44	0.72	0.26	0.65	0.34
Hill18 (18)	MAE	2.57	0.94	1.38	0.64	2.20	1.56
	MSE	-2.57	-0.31	-0.82	-0.54	-1.76	-1.56
	MAXE	3.91	1.85	5.01	1.67	3.95	3.83
	RMSE	2.68	1.07	1.69	0.83	2.34	1.84
	SD	0.78	1.06	1.52	0.65	1.58	1.00
X40x10 (220)	MAE	1.50	0.80	1.13	0.54	1.52	0.65
	MSE	-1.49	0.12	-1.06	-0.12	-1.51	-0.21
	MAXE	6.89	3.87	6.03	3.90	7.38	4.08
	RMSE	2.21	1.13	1.74	0.90	2.31	0.99
	SD	1.64	1.13	1.38	0.90	1.75	0.97

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
PNICO23 (23)	MAE	2.23	0.44	1.42	1.06	1.65	1.34
	MSE	2.23	0.38	1.25	1.05	1.54	1.34
	MAXE	4.95	1.26	4.49	3.92	4.61	4.25
	RMSE	2.50	0.57	2.00	1.49	2.12	1.81
	SD	1.16	0.43	1.60	1.08	1.49	1.25
CARBHB12 (12)	MAE	3.85	1.38	2.44	1.37	3.40	1.39
	MSE	3.85	0.14	2.44	1.07	3.40	1.08
	MAXE	7.41	2.78	4.86	3.51	6.48	3.42
	RMSE	4.58	1.59	2.98	1.79	3.96	1.85
	SD	2.58	1.66	1.79	1.50	2.11	1.56
ADIM6 (6)	MAE	1.11	1.01	0.21	1.75	0.74	2.06
	MSE	-1.11	1.01	-0.21	1.75	-0.74	2.06
	MAXE	1.65	1.67	0.32	2.87	1.02	3.48
	RMSE	1.20	1.11	0.21	1.93	0.78	2.27
	SD	0.52	0.51	0.07	0.88	0.29	1.05
HC12 (12)	MAE	1.35	1.04	0.45	1.58	0.88	1.80
	MSE	-1.35	0.33	-0.40	0.96	-0.88	1.23
	MAXE	2.37	1.70	1.21	2.78	1.96	3.69
	RMSE	1.44	1.13	0.54	1.70	0.96	2.01
	SD	0.50	1.13	0.37	1.47	0.41	1.67
HW30 (30)	MAE	1.68	0.52	1.12	0.35	1.64	0.50
	MSE	-1.68	-0.44	-1.12	-0.11	-1.64	-0.33
	MAXE	3.66	1.70	2.95	0.97	3.16	1.11
	RMSE	1.81	0.64	1.28	0.46	1.74	0.58
	SD	0.70	0.47	0.62	0.46	0.60	0.49
C2H4NT (75)	MAE	1.29	1.22	0.56	0.68	0.61	0.51
	MSE	-1.29	-1.22	0.45	0.23	-0.56	-0.43
	MAXE	2.92	4.23	1.83	1.62	1.22	1.79
	RMSE	1.51	1.56	0.65	0.78	0.71	0.67
	SD	0.79	0.98	0.47	0.75	0.44	0.52
CH4PAH (382)	MAE	0.66	0.50	0.36	0.35	0.42	0.32
	MSE	-0.65	-0.48	0.02	0.20	-0.33	0.27
	MAXE	2.15	4.48	2.62	1.36	1.34	2.72
	RMSE	0.79	0.88	0.46	0.42	0.50	0.61
	SD	0.46	0.74	0.46	0.37	0.37	0.54
CO2MOF (20)	MAE	2.43	0.59	1.64	0.53	2.36	0.50
	MSE	-2.43	-0.23	-1.64	-0.30	-2.36	-0.34
	MAXE	4.12	2.01	3.39	1.39	3.84	1.89
	RMSE	2.55	0.74	1.90	0.70	2.51	0.67
	SD	0.77	0.72	0.98	0.65	0.87	0.60
CO2PAH (249)	MAE	1.33	1.37	0.88	0.77	0.87	0.55
	MSE	-1.33	0.86	-0.33	0.73	-0.87	0.26
	MAXE	4.29	9.84	4.93	3.55	2.71	3.27
	RMSE	1.71	2.50	1.23	0.99	1.10	0.89
	SD	1.09	2.35	1.18	0.67	0.68	0.85
CO2NPHAC (96)	MAE	0.92	1.77	0.94	0.51	1.03	0.77
	MSE	-0.92	1.52	-0.67	0.50	-1.03	0.64
	MAXE	2.19	9.96	2.80	3.07	2.81	5.24
	RMSE	1.11	2.97	1.21	0.78	1.33	1.33
	SD	0.63	2.57	1.01	0.61	0.85	1.17
BzGas (129)	MAE	0.80	0.48	0.51	0.36	0.63	0.23
	MSE	-0.80	0.31	-0.41	-0.06	-0.62	0.04
	MAXE	2.40	3.66	2.43	1.31	1.50	1.07

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
	RMSE	0.95	0.88	0.77	0.48	0.73	0.29
	SD	0.51	0.83	0.65	0.48	0.37	0.29
Water38 (38)	MAE	36.41	1.39	26.22	4.81	36.38	3.39
	MSE	-36.41	0.77	-26.22	-4.81	-36.38	-3.39
	MAXE	71.87	2.88	50.33	8.44	71.53	6.24
	RMSE	39.56	1.60	28.29	5.13	39.46	3.71
	SD	15.67	1.42	10.77	1.82	15.47	1.53
Water1888 (1888)	MAE	2.05	1.31	1.74	0.41	2.16	0.46
	MSE	-2.05	0.63	-1.71	-0.14	-2.14	0.00
	MAXE	6.72	7.76	5.88	2.10	6.51	2.15
	RMSE	2.41	1.83	2.10	0.52	2.55	0.61
	SD	1.28	1.72	1.21	0.50	1.39	0.61
Water-2body (410)	MAE	0.87	0.34	0.68	0.16	0.93	0.13
	MSE	-0.85	0.10	-0.63	-0.09	-0.90	-0.06
	MAXE	4.92	2.28	3.69	1.30	4.61	1.46
	RMSE	1.59	0.59	1.25	0.24	1.75	0.21
	SD	1.34	0.59	1.08	0.23	1.51	0.21
B-set (160)	MAE	2.06	1.48	1.57	0.69	2.00	0.81
	MSE	-1.94	-0.84	-1.31	-0.06	-1.90	-0.49
	MAXE	7.48	12.19	7.49	3.84	7.71	3.32
	RMSE	2.80	2.30	2.26	0.94	2.83	1.17
	SD	2.02	2.15	1.85	0.94	2.10	1.06
F-set (160)	MAE	1.75	0.99	1.03	0.48	1.54	0.52
	MSE	-1.74	0.11	-0.85	0.30	-1.52	0.13
	MAXE	6.86	13.42	4.01	3.93	7.17	2.85
	RMSE	2.38	1.69	1.49	0.75	2.30	0.72
	SD	1.63	1.70	1.23	0.69	1.73	0.71
Si-set (152)	MAE	1.88	1.00	1.21	0.76	1.56	0.77
	MSE	-1.20	-0.65	-1.10	-0.36	-1.52	-0.53
	MAXE	6.46	6.80	5.41	6.38	5.21	5.02
	RMSE	2.54	1.61	1.83	1.38	2.13	1.30
	SD	2.25	1.47	1.46	1.34	1.50	1.19
P-set (120)	MAE	1.34	0.87	0.69	0.55	1.18	0.60
	MSE	-1.29	0.02	-0.56	0.31	-1.15	0.14
	MAXE	5.89	11.29	3.88	3.90	7.02	2.68
	RMSE	1.86	1.59	1.10	0.87	1.85	0.89
	SD	1.35	1.59	0.95	0.81	1.46	0.88
S-set (144)	MAE	1.27	0.44	0.61	0.45	1.00	0.46
	MSE	-1.27	-0.18	-0.48	0.08	-0.99	-0.13
	MAXE	4.03	2.98	2.41	2.07	3.26	3.10
	RMSE	1.61	0.64	0.84	0.60	1.29	0.66
	SD	0.99	0.62	0.69	0.60	0.82	0.65
Cl-set (160)	MAE	1.36	0.80	0.69	0.61	0.98	0.52
	MSE	-1.34	-0.02	-0.20	0.38	-0.93	-0.07
	MAXE	6.14	14.09	4.12	5.22	6.40	2.71
	RMSE	1.97	1.88	1.13	1.00	1.63	0.80
	SD	1.45	1.89	1.11	0.93	1.35	0.80
SSI-anionic (575)	MAE	5.54	2.52	3.37	1.96	4.77	2.38
	MSE	-5.53	-1.91	-3.23	-1.52	-4.74	-2.03
	MAXE	13.92	8.07	9.76	6.28	13.06	7.42
	RMSE	6.70	3.35	4.31	2.55	5.95	3.10
	SD	3.79	2.75	2.85	2.05	3.60	2.35
WatAA-anionic (64)	MAE	4.68	0.45	3.67	0.73	5.41	0.84

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
	MSE	-4.68	0.29	-3.67	-0.60	-5.41	-0.75
	MAXE	8.18	1.58	6.86	2.25	8.53	2.29
	RMSE	4.77	0.54	3.77	0.82	5.49	0.95
	SD	0.92	0.46	0.87	0.57	0.92	0.59
HSG-anionic (4)	MAE	6.18	1.69	4.56	1.31	6.07	1.46
	MSE	-6.18	-0.82	-4.56	-1.31	-6.07	-1.46
	MAXE	8.53	2.47	7.84	2.23	8.91	2.41
	RMSE	6.56	1.76	5.17	1.52	6.63	1.64
	SD	2.53	1.80	2.79	0.90	3.07	0.86
PLF547-anionic (155)	MAE	3.22	1.22	1.70	1.13	2.62	1.00
	MSE	-3.17	-0.88	-1.21	-0.04	-2.57	-0.62
	MAXE	19.38	16.95	12.06	6.08	15.83	6.69
	RMSE	5.18	2.60	3.14	1.69	4.47	1.87
	SD	4.11	2.45	2.90	1.70	3.66	1.77
IonicHB-anionic (24)	MAE	6.09	2.40	4.62	2.44	5.93	2.71
	MSE	-6.09	-2.39	-4.62	-2.43	-5.93	-2.71
	MAXE	8.49	5.32	7.95	3.58	9.58	4.07
	RMSE	6.59	2.76	5.16	2.63	6.52	2.90
	SD	2.57	1.41	2.34	1.03	2.77	1.05
IHB100x10-anionic (650)	MAE	8.17	4.32	5.92	4.44	7.32	4.82
	MSE	-8.16	-4.08	-5.91	-4.42	-7.31	-4.81
	MAXE	25.34	18.53	22.15	20.19	23.33	20.49
	RMSE	9.43	5.63	7.08	5.76	8.48	6.08
	SD	4.73	3.88	3.91	3.68	4.30	3.73
Ionic43-anionic (37)	MAE	8.61	4.90	6.34	4.78	7.80	4.60
	MSE	-8.61	-4.82	-6.34	-4.78	-7.80	-4.53
	MAXE	23.57	19.69	20.99	18.99	23.04	19.85
	RMSE	10.28	6.75	8.16	6.73	9.63	6.79
	SD	5.69	4.80	5.21	4.80	5.73	5.13
PEPCONF-Dipeptide (875)	MAE	1.29	0.89	1.02	0.56	1.10	0.58
	MSE	0.63	0.05	0.55	0.08	0.61	0.11
	MAXE	7.05	3.94	4.78	2.53	6.47	2.77
	RMSE	1.63	1.19	1.30	0.72	1.40	0.74
	SD	1.51	1.19	1.18	0.72	1.26	0.73
TPCONF (8)	MAE	2.25	0.76	2.05	0.57	1.71	0.27
	MSE	-2.00	0.18	-1.27	0.03	-0.98	0.08
	MAXE	4.53	1.53	4.34	1.11	3.05	0.74
	RMSE	2.92	0.83	2.48	0.72	1.99	0.34
	SD	2.28	0.86	2.27	0.77	1.85	0.35
P76 (71)	MAE	1.17	0.66	0.81	1.01	1.02	0.77
	MSE	1.01	-0.12	0.41	-0.19	0.98	-0.34
	MAXE	3.41	2.59	3.09	2.82	2.64	2.25
	RMSE	1.48	0.84	1.16	1.25	1.25	0.97
	SD	1.10	0.83	1.09	1.25	0.77	0.91
YMPJ (495)	MAE	1.08	0.85	0.72	0.62	1.06	0.52
	MSE	0.56	0.03	0.21	-0.44	0.92	-0.25
	MAXE	4.92	3.29	2.86	3.05	3.97	2.64
	RMSE	1.43	1.06	0.92	0.80	1.42	0.71
	SD	1.31	1.06	0.89	0.66	1.08	0.67
SPS (17)	MAE	0.52	1.16	0.69	0.38	0.53	0.49
	MSE	0.31	-0.86	0.57	0.30	0.33	-0.18
	MAXE	1.47	2.78	1.32	1.30	1.21	1.24
	RMSE	0.63	1.42	0.76	0.52	0.63	0.62

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
	SD	0.57	1.16	0.53	0.44	0.56	0.61
rSPS (45)	MAE	0.81	1.13	0.52	0.45	0.68	0.62
	MSE	0.69	0.70	0.27	-0.23	0.57	0.18
	MAXE	1.89	3.70	1.85	1.37	2.17	2.65
	RMSE	0.97	1.48	0.61	0.57	0.86	0.79
	SD	0.69	1.32	0.56	0.53	0.65	0.78
UpU46 (45)	MAE	1.42	1.13	0.97	0.94	1.32	0.97
	MSE	-1.03	-0.41	-0.34	0.29	-0.42	-0.02
	MAXE	6.42	2.97	3.24	3.87	5.45	3.61
	RMSE	1.87	1.38	1.25	1.17	1.63	1.23
	SD	1.58	1.34	1.22	1.14	1.59	1.24
SCONF (17)	MAE	4.03	0.98	2.47	0.63	3.29	0.75
	MSE	1.60	-0.65	0.94	-0.27	1.38	-0.36
	MAXE	10.39	2.25	6.25	3.21	7.94	2.51
	RMSE	4.67	1.22	2.86	1.05	3.75	1.03
	SD	4.52	1.07	2.78	1.04	3.60	0.99
DSCONF (27)	MAE	4.33	1.48	3.31	1.23	4.08	1.20
	MSE	3.26	1.00	2.88	0.44	3.46	0.42
	MAXE	11.13	3.53	8.84	3.57	10.31	3.80
	RMSE	5.65	1.78	4.17	1.52	5.13	1.55
	SD	4.71	1.51	3.07	1.48	3.87	1.52
SacchCONF (56)	MAE	2.15	1.41	1.24	0.69	1.28	0.86
	MSE	0.03	0.89	0.88	0.14	0.79	0.58
	MAXE	7.64	5.38	4.11	1.92	3.46	2.35
	RMSE	2.81	1.93	1.50	0.82	1.57	1.04
	SD	2.84	1.73	1.22	0.81	1.37	0.87
CCONF (426)	MAE	3.32	2.82	1.50	0.78	1.94	0.93
	MSE	-1.61	2.70	0.09	-0.23	-0.42	0.28
	MAXE	16.20	8.83	6.27	3.08	8.64	3.88
	RMSE	4.40	3.44	1.90	1.00	2.50	1.17
	SD	4.10	2.14	1.90	0.98	2.47	1.13
ACONF (15)	MAE	0.27	0.08	0.33	0.17	0.04	0.39
	MSE	-0.27	-0.02	-0.33	-0.17	-0.02	0.39
	MAXE	0.63	0.18	0.66	0.32	0.07	0.89
	RMSE	0.31	0.09	0.36	0.19	0.04	0.45
	SD	0.16	0.10	0.17	0.08	0.04	0.23
BCONF (64)	MAE	2.55	0.85	1.75	0.45	2.44	0.39
	MSE	2.51	-0.74	1.66	0.34	2.42	0.02
	MAXE	4.30	2.03	2.80	1.29	3.61	1.01
	RMSE	2.70	0.97	1.83	0.56	2.56	0.47
	SD	1.01	0.64	0.79	0.44	0.84	0.48
PentCONF (342)	MAE	0.30	0.21	0.13	0.17	0.06	0.42
	MSE	-0.30	0.16	-0.05	-0.16	0.00	0.42
	MAXE	1.06	1.09	0.50	0.51	0.27	0.81
	RMSE	0.39	0.31	0.17	0.20	0.08	0.46
	SD	0.25	0.27	0.16	0.12	0.08	0.18
Undecamer125 (124)	MAE	2.27	1.15	1.78	0.60	1.79	0.50
	MSE	1.38	0.83	0.85	-0.58	1.19	-0.44
	MAXE	4.89	2.49	4.76	2.21	3.64	1.73
	RMSE	2.59	1.37	2.04	0.75	2.06	0.65
	SD	2.20	1.09	1.86	0.48	1.69	0.48
ICONF (17)	MAE	0.51	1.07	0.67	0.54	0.57	0.74
	MSE	0.17	-0.04	0.16	0.08	0.34	0.53

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
	MAXE	2.66	3.45	1.63	2.31	1.67	1.71
	RMSE	0.81	1.42	0.85	0.81	0.73	0.94
	SD	0.82	1.47	0.86	0.83	0.66	0.81
MCONF (51)	MAE	1.14	0.83	0.78	0.44	0.89	0.22
	MSE	0.99	-0.65	0.75	0.21	0.83	-0.01
	MAXE	2.31	2.15	1.43	1.20	1.41	0.55
	RMSE	1.31	0.96	0.85	0.55	0.95	0.27
	SD	0.87	0.72	0.40	0.51	0.47	0.27
Torsion21 (189)	MAE	0.39	0.57	0.84	0.86	0.58	0.69
	MSE	0.12	0.47	0.65	0.54	0.46	0.49
	MAXE	1.87	1.85	2.93	2.34	2.15	2.39
	RMSE	0.56	0.73	1.07	1.07	0.74	0.88
	SD	0.54	0.56	0.84	0.92	0.58	0.73
37Conf8 (258)	MAE	1.43	1.05	1.02	0.78	1.08	0.82
	MSE	0.19	-0.51	0.19	-0.40	0.51	-0.20
	MAXE	10.35	5.93	7.43	4.31	9.13	3.50
	RMSE	2.01	1.36	1.49	1.03	1.63	1.04
	SD	2.00	1.27	1.48	0.95	1.55	1.02
DCONF (2142)	MAE	0.67	0.62	0.49	0.36	0.51	0.41
	MSE	0.50	0.46	0.41	0.16	0.44	0.26
	MAXE	3.30	3.77	2.35	1.67	2.21	2.11
	RMSE	0.93	0.90	0.70	0.51	0.70	0.59
	SD	0.79	0.78	0.56	0.48	0.55	0.53
MolCONF (5623)	MAE	0.50	0.51	0.39	0.34	0.41	0.36
	MSE	0.07	0.04	-0.04	-0.10	0.01	-0.07
	MAXE	10.64	6.66	4.63	6.26	6.37	6.41
	RMSE	0.91	0.87	0.63	0.55	0.73	0.60
	SD	0.91	0.86	0.63	0.54	0.73	0.60
PEPCONF-Dipeptide-anionic (175)	MAE	1.23	0.94	0.94	0.77	1.03	0.87
	MSE	-0.07	-0.19	0.05	-0.17	0.07	-0.03
	MAXE	4.53	4.19	3.03	2.66	3.80	2.59
	RMSE	1.59	1.21	1.21	0.97	1.33	1.07
	SD	1.60	1.20	1.21	0.96	1.33	1.07
MolCONF-anionic (79)	MAE	0.70	0.79	0.18	0.47	0.25	0.42
	MSE	0.41	0.53	0.05	0.05	0.08	0.01
	MAXE	4.50	4.12	0.75	2.94	1.06	2.25
	RMSE	1.48	1.41	0.27	0.84	0.40	0.70
	SD	1.43	1.31	0.26	0.84	0.39	0.71
PAH6 (6)	MAE	13.79	13.71	16.90	16.12	16.96	15.56
	MSE	-13.79	-13.71	-16.90	-16.12	-16.96	-15.56
	MAXE	34.80	34.60	41.28	40.20	41.27	39.46
	RMSE	17.95	17.86	21.55	20.80	21.59	20.29
	SD	12.58	12.54	14.65	14.40	14.62	14.26
IDISP (6)	MAE	3.97	2.92	1.11	1.75	2.53	3.79
	MSE	3.97	1.99	0.18	-0.88	0.14	-1.84
	MAXE	10.34	9.33	3.41	3.46	4.16	8.00
	RMSE	5.28	4.21	1.53	2.06	2.91	4.82
	SD	3.81	4.07	1.66	2.04	3.18	4.88
AlkIsomer11 (11)	MAE	0.56	0.69	0.11	0.43	1.01	0.34
	MSE	0.56	0.69	0.09	-0.43	1.01	0.33
	MAXE	1.35	2.09	0.25	1.05	2.74	1.21
	RMSE	0.67	0.88	0.12	0.52	1.22	0.47
	SD	0.39	0.58	0.09	0.29	0.72	0.34

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
Styrene45 (44)	MAE	5.46	3.50	2.89	2.07	3.28	2.05
	MSE	3.63	1.01	-0.15	-0.80	2.69	0.59
	MAXE	16.29	10.56	13.84	6.76	10.41	5.58
	RMSE	7.01	4.25	3.91	2.54	3.92	2.45
	SD	6.07	4.18	3.96	2.44	2.88	2.41
H2O16Rel5 (4)	MAE	5.39	3.42	4.82	0.84	4.47	0.84
	MSE	5.11	3.42	4.51	0.84	4.15	0.84
	MAXE	7.71	4.47	6.96	1.50	6.37	1.38
	RMSE	6.08	3.73	5.41	0.97	5.00	0.93
	SD	3.81	1.72	3.46	0.58	3.22	0.45
H2O20Rel10 (9)	MAE	2.18	1.86	1.73	0.94	1.71	1.04
	MSE	-1.82	-1.74	-1.71	-0.41	-1.20	-0.43
	MAXE	10.05	6.60	8.45	2.50	7.91	2.82
	RMSE	3.70	2.60	3.07	1.19	2.87	1.30
	SD	3.42	2.05	2.71	1.18	2.77	1.30
SW49Rel6 (17)	MAE	5.19	1.92	2.36	0.86	3.81	1.68
	MSE	5.16	1.92	2.34	0.85	3.76	1.67
	MAXE	19.60	3.34	9.73	3.34	15.78	5.54
	RMSE	6.85	2.31	3.22	1.12	5.34	2.04
	SD	4.64	1.33	2.27	0.75	3.91	1.21
SW49Rel345 (28)	MAE	3.96	1.03	1.83	0.49	3.09	1.01
	MSE	-0.28	0.10	-0.14	0.01	-0.17	0.02
	MAXE	9.37	1.84	4.67	1.60	7.55	2.74
	RMSE	4.61	1.19	2.14	0.62	3.63	1.27
	SD	4.69	1.21	2.18	0.63	3.70	1.30
TAUT15 (15)	MAE	2.16	1.49	1.10	1.82	1.56	1.70
	MSE	-0.90	-0.62	-0.41	0.94	-0.56	0.71
	MAXE	5.92	4.56	3.37	4.38	5.28	4.42
	RMSE	2.61	1.83	1.45	2.20	1.93	1.96
	SD	2.53	1.79	1.44	2.06	1.91	1.90
PArel (20)	MAE	2.66	2.10	2.39	1.42	2.35	1.53
	MSE	-0.45	0.94	0.17	0.53	-0.02	0.71
	MAXE	11.15	8.83	9.61	3.92	9.06	4.86
	RMSE	3.87	2.99	3.53	1.84	3.37	2.02
	SD	3.94	2.92	3.62	1.81	3.46	1.94
EIE22 (22)	MAE	2.90	2.62	1.03	0.97	1.63	1.54
	MSE	2.85	2.62	0.96	0.88	1.58	1.54
	MAXE	5.91	5.10	2.30	1.85	3.53	2.91
	RMSE	3.20	2.88	1.24	1.09	1.89	1.71
	SD	1.48	1.21	0.80	0.65	1.07	0.75
ISO34 (34)	MAE	3.78	1.73	2.75	1.30	2.87	1.49
	MSE	-1.67	-0.68	-1.58	-0.93	-1.33	-0.72
	MAXE	20.99	5.85	9.98	5.62	13.46	6.94
	RMSE	5.48	2.17	3.54	1.82	3.90	2.10
	SD	5.30	2.10	3.22	1.59	3.72	2.00
ISOL24 (24)	MAE	6.89	2.83	3.20	1.81	2.55	2.63
	MSE	-3.33	0.17	-0.49	-0.53	-0.42	0.20
	MAXE	22.58	8.60	11.87	5.31	6.80	10.74
	RMSE	8.86	3.82	4.14	2.32	3.33	3.70
	SD	8.38	3.89	4.20	2.31	3.38	3.77
BH9 (898)	MAE	12.72	7.40	3.43	2.98	4.16	2.88
	MSE	-12.27	-6.64	-1.71	-1.56	-2.39	-1.56
	MAXE	96.05	78.89	89.84	91.90	87.14	92.36

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
	RMSE	15.25	10.09	6.98	6.57	7.57	6.55
	SD	9.06	7.61	6.77	6.38	7.19	6.37
E2SN2 (418)	MAE	7.33	3.57	4.59	3.88	4.01	2.89
	MSE	-7.17	-0.84	-2.10	-2.25	-1.98	-1.46
	MAXE	18.98	14.11	19.45	18.74	17.28	14.77
	RMSE	8.32	4.55	5.71	5.11	4.97	3.93
	SD	4.23	4.47	5.32	4.59	4.56	3.65
BHPERI26 (26)	MAE	7.24	4.47	2.91	2.99	2.25	2.60
	MSE	-7.24	-4.21	-1.97	-2.67	-1.17	-2.14
	MAXE	12.15	9.67	6.44	5.48	5.90	4.70
	RMSE	7.66	5.17	3.30	3.50	2.82	3.00
	SD	2.57	3.05	2.70	2.30	2.61	2.14
CRBH20 (20)	MAE	13.52	1.34	3.78	1.06	2.31	1.61
	MSE	-13.52	-0.08	3.78	0.04	2.03	-1.26
	MAXE	16.42	3.47	6.75	4.29	4.29	4.05
	RMSE	13.59	1.73	4.47	1.53	2.66	2.01
	SD	1.33	1.77	2.44	1.57	1.76	1.60
DBH24 (24)	MAE	10.99	5.23	4.18	3.27	6.00	4.45
	MSE	-10.36	-3.90	-1.21	-1.14	-3.97	-3.00
	MAXE	33.88	20.20	18.30	15.53	22.72	16.60
	RMSE	13.84	7.16	6.54	5.01	8.23	6.01
	SD	9.37	6.13	6.57	4.98	7.37	5.32
HTBH38 (38)	MAE	10.01	5.09	2.61	1.93	4.53	2.58
	MSE	-10.01	-4.93	-1.41	-0.47	-4.31	-2.13
	MAXE	21.52	15.16	9.21	5.08	15.97	7.46
	RMSE	11.14	6.10	3.42	2.40	5.67	3.10
	SD	4.96	3.63	3.16	2.39	3.74	2.28
NHTBH38 (38)	MAE	13.59	5.54	5.68	4.12	8.14	5.37
	MSE	-12.91	-3.40	-2.43	-1.65	-5.55	-3.04
	MAXE	41.16	21.28	25.28	15.54	30.16	19.01
	RMSE	17.22	7.52	8.37	5.86	10.85	7.10
	SD	11.55	6.80	8.12	5.70	9.44	6.50
PX13 (13)	MAE	33.76	2.17	23.05	3.65	29.49	3.59
	MSE	-33.76	-0.25	-23.05	-3.56	-29.49	-3.35
	MAXE	60.70	5.12	43.79	5.97	58.58	5.41
	RMSE	36.57	2.73	25.74	3.96	33.23	3.76
	SD	14.63	2.83	11.94	1.80	15.94	1.79
WCPT27 (27)	MAE	12.81	3.18	7.82	2.37	9.66	2.49
	MSE	-12.81	1.11	-5.54	0.81	-7.31	1.36
	MAXE	28.78	7.17	15.89	9.91	19.22	9.52
	RMSE	15.19	3.77	9.10	3.20	11.36	3.36
	SD	8.31	3.67	7.36	3.15	8.86	3.13
INV24 (24)	MAE	2.28	3.48	2.04	3.31	1.73	4.08
	MSE	-1.26	-2.17	0.80	-1.93	0.25	-1.33
	MAXE	8.17	10.82	11.57	16.90	8.48	11.56
	RMSE	3.08	4.70	3.21	5.10	2.42	5.33
	SD	2.87	4.26	3.17	4.82	2.46	5.27
BHROT27 (27)	MAE	0.65	0.90	0.55	0.65	0.62	0.73
	MSE	0.44	0.86	0.53	0.34	0.59	0.62
	MAXE	1.80	2.98	1.81	2.08	1.81	2.81
	RMSE	0.84	1.25	0.74	0.91	0.84	1.07
	SD	0.73	0.92	0.53	0.86	0.62	0.88
	MAE	7.05	3.42	4.05	2.86	3.93	2.55

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
Grambow2020-ωB97xD3 (23922)	MSE	-6.61	-1.74	3.26	0.72	3.10	0.75
	MAXE	38.51	33.49	40.04	32.16	39.24	30.32
	RMSE	8.67	4.48	5.39	4.03	4.99	3.60
	SD	5.61	4.13	4.30	3.97	3.90	3.52
Grambow2020-B97D3(32722)	MAE	8.01	4.08	5.14	3.76	4.78	3.38
	MSE	-7.46	-2.31	4.32	1.43	3.88	1.32
	MAXE	193.94	190.51	212.07	212.66	210.23	210.74
	RMSE	9.86	5.67	7.59	6.22	6.96	5.75
	SD	6.44	5.17	6.24	6.06	5.78	5.60
MN-RE (7555)	MAE	9.06	6.52	8.87	6.09	9.04	5.94
	MSE	2.29	1.04	1.30	1.11	1.95	1.36
	MAXE	61.18	53.95	80.84	53.88	82.57	52.34
	RMSE	12.03	8.81	13.02	8.74	13.50	8.33
	SD	11.81	8.75	12.96	8.67	13.36	8.22
BH9-RE (449)	MAE	5.39	3.12	3.40	2.02	3.49	2.43
	MSE	1.79	-1.02	-1.41	-0.64	-1.56	-1.21
	MAXE	42.25	38.22	38.24	39.52	37.53	38.04
	RMSE	7.31	4.84	5.00	3.82	5.23	4.08
	SD	7.10	4.74	4.81	3.77	5.00	3.90
WCPT6 (6)	MAE	3.71	1.37	2.74	2.42	3.42	2.22
	MSE	3.04	0.84	2.74	1.66	3.12	1.77
	MAXE	6.24	2.08	4.37	4.50	6.01	4.47
	RMSE	3.95	1.51	3.06	2.74	3.75	2.64
	SD	2.77	1.38	1.49	2.39	2.28	2.15
NBPRC (6)	MAE	2.81	1.77	2.13	3.34	1.73	1.84
	MSE	2.67	0.53	1.98	2.95	0.83	1.73
	MAXE	7.56	2.67	3.66	7.99	3.09	5.34
	RMSE	3.64	1.93	2.44	4.08	2.00	2.51
	SD	2.71	2.03	1.56	3.09	1.99	2.00
PlatonicHD6 (6)	MAE	3.90	12.32	7.73	1.63	3.77	4.04
	MSE	-3.65	-12.32	5.93	0.21	3.44	-1.19
	MAXE	8.19	17.60	17.41	3.28	6.02	7.85
	RMSE	4.78	13.45	9.11	1.91	4.17	4.52
	SD	3.38	5.90	7.58	2.07	2.57	4.77
PlatonicID6 (6)	MAE	7.15	11.88	14.99	1.89	17.34	3.10
	MSE	7.15	-11.88	14.30	-1.46	17.34	2.33
	MAXE	13.71	17.16	34.15	2.76	26.81	5.77
	RMSE	8.49	12.98	18.11	2.04	18.23	3.91
	SD	5.00	5.73	12.17	1.56	6.15	3.43
PlatonicIG6 (6)	MAE	5.87	9.23	10.82	4.81	26.80	15.51
	MSE	-5.87	-9.23	9.49	4.75	26.80	15.51
	MAXE	12.63	14.51	24.52	11.19	45.72	25.55
	RMSE	7.03	10.23	12.86	6.16	28.75	16.58
	SD	4.23	4.83	9.51	4.29	11.39	6.42
DC13 (12)	MAE	10.85	5.70	10.34	6.04	9.24	5.91
	MSE	5.66	1.00	-2.52	-3.19	-0.98	-2.40
	MAXE	27.88	11.09	24.49	20.09	30.50	17.53
	RMSE	13.33	6.84	13.20	8.21	12.43	7.93
	SD	12.60	7.06	13.53	7.90	12.94	7.89
DARC (14)	MAE	4.85	2.50	4.99	2.44	5.39	3.78
	MSE	4.58	-2.03	-4.99	-2.44	-5.39	-3.78
	MAXE	7.53	4.96	8.43	4.14	11.51	5.85
	RMSE	5.46	3.08	5.38	2.81	5.94	4.11

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
	SD	3.09	2.41	2.09	1.46	2.58	1.68
AlkIsod14 (14)	MAE	1.68	0.33	1.75	0.22	2.53	0.72
	MSE	-1.68	-0.02	-1.75	-0.09	-2.53	-0.72
	MAXE	2.92	0.81	3.01	0.62	4.89	1.48
	RMSE	1.81	0.40	1.90	0.28	2.77	0.79
	SD	0.72	0.42	0.76	0.28	1.17	0.34
CR20 (20)	MAE	2.46	2.70	3.92	2.61	4.52	3.34
	MSE	-1.93	2.70	3.92	2.28	4.52	3.27
	MAXE	3.84	7.10	9.42	4.04	9.99	5.05
	RMSE	2.70	3.18	4.64	2.77	5.15	3.55
	SD	1.93	1.72	2.56	1.62	2.53	1.39
RC21 (21)	MAE	8.00	7.18	5.57	2.96	7.15	4.09
	MSE	6.98	5.96	5.24	1.91	7.00	3.85
	MAXE	18.89	15.25	12.76	5.99	17.31	8.03
	RMSE	9.46	8.52	6.30	3.37	8.27	4.70
	SD	6.53	6.23	3.59	2.85	4.51	2.77
G2RC (23)	MAE	13.78	4.97	8.19	3.11	9.60	4.43
	MSE	6.08	0.20	-0.54	-0.11	-0.76	-0.66
	MAXE	42.11	16.73	21.83	7.17	29.37	10.95
	RMSE	17.21	6.50	10.41	3.66	13.09	5.68
	SD	16.47	6.64	10.63	3.74	13.37	5.77
BH76RC (30)	MAE	8.72	4.02	7.35	3.98	8.44	4.43
	MSE	0.21	-0.94	-1.06	-1.85	-1.33	-2.21
	MAXE	46.48	20.72	42.05	21.73	45.54	23.84
	RMSE	13.07	6.12	11.07	5.74	12.12	6.43
	SD	13.29	6.16	11.20	5.52	12.25	6.14
BSR36 (36)	MAE	2.74	2.61	3.62	0.77	4.81	1.03
	MSE	-2.74	2.55	-3.62	0.77	-4.81	0.17
	MAXE	7.01	12.32	12.62	3.15	13.06	5.07
	RMSE	3.15	3.70	4.53	0.99	5.58	1.45
	SD	1.58	2.72	2.76	0.62	2.86	1.46
FH51 (51)	MAE	7.73	3.85	5.62	2.86	6.96	3.78
	MSE	4.12	1.00	0.24	1.22	0.21	0.85
	MAXE	29.45	19.12	20.83	19.86	23.35	20.72
	RMSE	10.86	5.50	8.01	4.82	9.50	5.52
	SD	10.15	5.46	8.09	4.71	9.59	5.51
DIE60 (60)	MAE	2.39	1.12	1.16	0.67	1.29	1.05
	MSE	2.39	1.05	1.11	0.46	1.24	0.85
	MAXE	4.76	2.52	3.07	1.92	3.14	3.05
	RMSE	2.55	1.35	1.44	0.89	1.56	1.43
	SD	0.90	0.85	0.92	0.76	0.95	1.16
PlatonicTAE6 (6)	MAE	3.72	20.28	13.01	26.07	17.30	26.86
	MSE	-3.72	20.28	13.01	-26.07	17.30	-26.86
	MAXE	9.77	32.99	20.16	54.13	42.17	48.56
	RMSE	5.04	22.04	13.52	30.05	21.29	29.25
	SD	3.73	9.46	4.05	16.37	13.59	12.68
AlkAtom19 (19)	MAE	2.62	7.27	2.95	2.82	16.04	2.49
	MSE	2.62	-7.27	-2.74	2.46	-16.04	-2.49
	MAXE	4.62	11.08	5.09	5.60	23.95	4.07
	RMSE	2.78	7.65	3.25	3.19	16.90	2.62
	SD	0.96	2.47	1.80	2.08	5.46	0.86
TAE-W4-17 (194)	MAE	8.94	8.87	8.13	7.06	8.18	6.33
	MSE	3.67	8.02	-6.54	-3.75	-4.46	-1.19

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
	MAXE	58.07	45.96	76.36	35.76	73.67	45.05
	RMSE	12.85	11.33	13.11	9.19	13.41	8.84
	SD	12.35	8.03	11.39	8.41	12.68	8.79
MOLdef (9298)	MAE	1.39	1.22	0.52	0.64	0.55	0.64
	MSE	-0.53	0.28	0.33	0.22	0.34	0.29
	MAXE	31.78	24.92	15.02	19.46	15.37	15.86
	RMSE	2.47	1.98	0.95	1.31	0.98	1.27
	SD	2.41	1.96	0.90	1.29	0.92	1.24
MOLdef-H2O (990)	MAE	0.90	0.53	0.39	0.16	0.43	0.31
	MSE	-0.30	-0.24	-0.03	0.09	-0.05	0.15
	MAXE	5.52	3.97	3.11	2.03	3.56	2.53
	RMSE	1.40	0.79	0.62	0.27	0.68	0.47
	SD	1.37	0.76	0.61	0.25	0.68	0.45
MOLdef-TS (6294)	MAE	4.76	3.69	1.81	1.80	1.70	1.70
	MSE	4.37	2.44	0.69	0.68	0.71	0.62
	MAXE	101.46	100.80	97.83	98.84	97.79	98.60
	RMSE	7.04	5.71	3.87	3.81	3.77	3.74
	SD	5.52	5.16	3.81	3.74	3.71	3.69
BSE49-expt (1969)	MAE	6.36	4.14	2.76	2.19	3.27	2.89
	MSE	5.96	3.23	-0.16	-0.37	2.84	1.79
	MAXE	26.03	41.84	22.06	18.52	26.34	91.18
	RMSE	7.40	5.30	3.61	3.02	4.36	4.84
	SD	4.37	4.20	3.60	3.00	3.31	4.50
BSE49-non-expt (2533)	MAE	6.35	3.18	3.41	2.40	3.95	2.25
	MSE	5.37	1.08	1.78	0.29	3.43	0.84
	MAXE	26.23	17.78	58.40	59.36	36.18	49.03
	RMSE	7.46	3.98	5.64	4.54	4.99	3.50
	SD	5.18	3.84	5.35	4.53	3.63	3.40

Table S4. Detailed error analysis with respect to reference data in the validation set. The numbers in bracket in the first column indicates the number of data points. The various shorthand notations are as follows: MAE = mean absolute error in kcal/mol, MSE = mean signed error in kcal/mol, MAXE = maximum absolute error in kcal/mol, RMSE = root-mean-square error in kcal/mol, and SD = standard deviation in kcal/mol.

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
BlindNCI (80)	MAE	0.99	0.58	0.94	0.31	1.05	0.39
	MSE	-0.96	0.27	-0.80	0.11	-1.04	0.18
	MAXE	5.40	6.72	7.42	2.10	5.92	4.57
	RMSE	1.73	1.30	1.78	0.52	1.82	0.84
	SD	1.45	1.28	1.61	0.51	1.51	0.82
DES15K (11474)	MAE	2.03	1.48	1.50	0.87	2.02	1.18
	MSE	-1.98	0.72	-1.33	0.58	-1.90	0.77
	MAXE	10.74	14.53	10.96	10.07	12.64	14.29
	RMSE	2.44	2.46	1.98	1.25	2.60	1.82
	SD	1.43	2.36	1.47	1.11	1.78	1.64
NENCI-2021 (5859)	MAE	2.14	1.27	1.45	0.90	2.23	1.10
	MSE	-2.02	0.61	-1.16	0.54	-1.97	0.55
	MAXE	8.38	14.77	6.92	12.02	8.59	19.88
	RMSE	2.50	2.07	1.92	1.56	2.77	2.12
	SD	1.48	1.98	1.53	1.46	1.94	2.05

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
R160x6 (960)	MAE	0.97	1.27	0.68	0.74	0.95	0.87
	MSE	-0.80	1.03	-0.22	0.45	-0.66	0.50
	MAXE	4.73	8.04	6.01	6.68	6.51	7.12
	RMSE	1.27	1.85	1.03	1.06	1.28	1.24
	SD	0.99	1.54	1.00	0.96	1.09	1.13
R739x5 (4330)	MAE	0.81	0.95	0.50	0.45	0.77	0.52
	MSE	-0.77	0.75	-0.29	0.28	-0.72	0.26
	MAXE	5.56	8.51	4.35	4.11	6.40	3.47
	RMSE	1.05	1.39	0.76	0.62	1.08	0.74
	SD	0.72	1.17	0.70	0.56	0.80	0.69
CE20 (20)	MAE	21.46	2.86	15.18	3.30	21.40	3.81
	MSE	21.46	0.88	15.18	3.30	21.40	3.81
	MAXE	45.15	9.01	29.85	5.98	45.16	8.49
	RMSE	24.74	4.08	17.37	3.70	24.69	4.56
	SD	12.62	4.09	8.67	1.72	12.63	2.57
CHAL336 (48)	MAE	2.19	1.27	1.17	1.10	1.71	1.60
	MSE	-1.96	-1.05	-0.97	-1.09	-1.52	-1.60
	MAXE	3.68	9.71	4.88	9.28	3.58	10.91
	RMSE	2.37	2.73	1.36	1.70	1.85	2.22
	SD	1.34	2.55	0.97	1.32	1.07	1.56
XB45 (33)	MAE	6.40	2.81	3.34	6.78	4.07	6.86
	MSE	6.40	2.31	3.21	6.71	4.01	6.76
	MAXE	14.58	11.27	8.47	22.70	8.97	23.53
	RMSE	7.47	4.46	3.96	9.72	4.71	9.94
	SD	3.92	3.87	2.36	7.14	2.50	7.39
WaterOrg (2376)	MAE	3.29	0.77	1.67	0.63	3.17	0.46
	MSE	-3.29	0.50	-1.67	0.61	-3.17	0.34
	MAXE	8.56	2.47	5.02	1.77	7.63	1.83
	RMSE	3.53	0.99	1.87	0.72	3.41	0.56
	SD	1.29	0.86	0.86	0.37	1.26	0.44
H2O20Bind10 (10)	MAE	145.88	24.40	107.24	15.19	154.83	13.15
	MSE	-145.88	24.40	-107.24	-15.19	-154.83	-13.15
	MAXE	154.29	26.52	114.15	17.32	161.66	15.59
	RMSE	145.91	24.47	107.27	15.22	154.85	13.20
	SD	3.27	2.01	2.61	1.12	2.64	1.23
L7 (7)	MAE	7.40	4.76	3.76	2.66	4.28	2.42
	MSE	-7.40	-3.45	1.24	1.98	-4.28	-0.50
	MAXE	10.95	8.46	6.17	10.86	11.51	6.71
	RMSE	7.91	5.36	4.17	4.32	5.54	3.05
	SD	3.03	4.43	4.30	4.16	3.79	3.25
S12L (10)	MAE	17.23	11.28	5.80	5.16	13.47	8.41
	MSE	-17.23	-10.87	-5.80	-4.39	-13.47	-8.25
	MAXE	25.94	33.28	9.61	16.89	20.28	26.51
	RMSE	18.08	16.26	6.13	7.88	14.19	12.62
	SD	5.76	12.74	2.07	6.90	4.71	10.07
S30L (26)	MAE	18.14	13.00	3.10	6.66	13.04	9.18
	MSE	-18.14	-11.38	-2.78	-1.89	-13.04	-7.06
	MAXE	34.15	41.65	12.18	16.21	34.84	30.15
	RMSE	19.56	18.21	4.24	8.08	14.33	12.98
	SD	7.48	14.50	3.27	8.02	6.05	11.11
C60dimer (14)	MAE	3.40	4.49	2.09	0.71	1.78	3.97
	MSE	-3.40	-4.49	2.08	0.05	-1.78	-3.97
	MAXE	7.17	9.74	4.61	1.62	3.12	9.29

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
	RMSE	3.78	5.10	2.36	0.88	1.99	4.57
	SD	1.73	2.51	1.16	0.91	0.92	2.35
Ni2021 (11)	MAE	27.30	13.61	8.33	22.48	27.07	13.30
	MSE	-26.04	12.91	-2.11	22.48	-25.44	13.16
	MAXE	81.62	57.74	13.91	80.46	78.22	53.41
	RMSE	33.99	19.75	9.62	30.79	33.23	18.87
	SD	22.91	15.67	9.85	22.06	22.42	14.17
HW6Cl-anionic (6)	MAE	21.26	5.32	16.80	5.92	20.79	5.12
	MSE	-21.26	-5.32	-16.80	-5.92	-20.79	-5.12
	MAXE	39.46	8.37	29.18	9.04	38.89	8.15
	RMSE	24.77	5.93	19.26	6.56	24.25	5.77
	SD	13.92	2.86	10.31	3.08	13.66	2.93
HW6F-anionic (6)	MAE	58.73	33.74	47.16	30.73	56.76	32.90
	MSE	-58.73	-33.74	-47.16	-30.73	-56.76	-32.90
	MAXE	87.17	41.57	68.21	39.25	84.79	42.02
	RMSE	62.62	35.13	50.01	31.76	60.53	34.06
	SD	23.80	10.70	18.22	8.80	23.04	9.64
FmH2O10-anionic (10)	MAE	126.92	41.86	99.41	47.18	125.46	48.42
	MSE	-126.92	-41.86	-99.41	-47.18	-125.46	-48.42
	MAXE	128.35	45.12	100.71	47.86	127.16	49.41
	RMSE	126.93	41.90	99.42	47.18	125.47	48.43
	SD	1.68	1.94	1.47	0.53	1.53	0.67
SW49Bind345-anionic (30)	MAE	13.38	2.50	8.73	2.18	12.04	2.60
	MSE	-13.38	-2.50	-8.71	-2.17	-12.03	-2.60
	MAXE	24.84	5.33	16.15	4.05	22.72	4.91
	RMSE	15.90	2.88	10.44	2.57	14.40	3.02
	SD	8.74	1.46	5.84	1.40	8.06	1.55
SW49Bind6-anionic (18)	MAE	30.84	4.61	20.83	5.03	28.66	5.36
	MSE	-30.84	-4.61	-20.83	-5.03	-28.66	-5.36
	MAXE	35.90	6.43	23.16	5.95	32.41	7.05
	RMSE	31.17	4.79	20.95	5.08	28.91	5.49
	SD	4.66	1.36	2.28	0.75	3.89	1.23
Anionpi-anionic (16)	MAE	11.89	7.72	7.89	5.11	8.21	5.58
	MSE	-11.28	-7.09	-7.63	-4.12	-7.78	-3.83
	MAXE	29.83	20.84	23.82	15.59	24.00	16.16
	RMSE	16.08	10.74	10.71	7.12	11.37	7.56
	SD	11.84	8.33	7.76	6.00	8.56	6.73
IL236-anionic (236)	MAE	9.70	4.50	6.31	2.86	7.99	2.98
	MSE	-9.70	-4.47	-6.31	-2.80	-7.99	-2.93
	MAXE	16.78	10.38	12.07	6.74	14.85	7.24
	RMSE	10.09	5.02	6.71	3.31	8.52	3.45
	SD	2.79	2.29	2.27	1.76	2.96	1.81
DES15K-anionic (1281)	MAE	6.73	3.89	5.40	4.34	6.34	4.18
	MSE	-6.63	-2.70	-5.38	-4.27	-6.29	-4.07
	MAXE	46.04	39.88	38.09	35.37	43.60	36.56
	RMSE	10.19	6.73	8.22	6.75	9.52	6.88
	SD	7.73	6.16	6.22	5.23	7.14	5.54
NENCI-2021-anionic (889)	MAE	11.64	7.12	7.08	4.15	8.71	4.92
	MSE	-10.83	-5.99	-5.93	-3.28	-7.25	-3.41
	MAXE	41.36	24.91	37.45	19.90	40.92	20.60
	RMSE	16.31	10.01	9.90	5.93	11.96	6.73
	SD	12.20	8.03	7.94	4.94	9.52	5.80
	MAE	16.95	10.82	10.59	11.27	12.13	13.11

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
CHAL336-anionic (19)	MSE	-16.95	-10.82	-10.58	-11.27	-12.00	-13.11
	MAXE	40.20	23.96	32.20	30.32	36.18	34.04
	RMSE	22.94	14.09	14.97	14.55	17.56	17.17
	SD	15.88	9.28	10.89	9.46	13.16	11.39
XB45-anionic (12)	MAE	26.99	21.86	22.01	23.06	23.38	24.01
	MSE	26.99	21.86	22.01	23.06	23.38	24.01
	MAXE	73.46	56.63	52.90	47.01	62.09	52.22
	RMSE	35.00	28.57	28.91	28.25	31.40	29.93
	SD	23.28	19.21	19.57	17.04	21.90	18.67
S30L-anionic (2)	MAE	20.66	5.14	14.22	3.40	19.24	2.73
	MSE	-20.66	-5.14	-14.22	-3.40	-19.24	-2.73
	MAXE	21.66	5.54	16.83	5.07	20.30	3.15
	RMSE	20.69	5.15	14.46	3.79	19.27	2.76
	SD	1.41	0.56	3.69	2.36	1.50	0.59
SafroleCONF (5)	MAE	0.53	0.56	0.49	0.50	0.47	0.49
	MSE	-0.53	-0.56	-0.49	-0.50	-0.47	-0.49
	MAXE	1.13	1.26	1.07	1.10	1.08	1.08
	RMSE	0.71	0.78	0.68	0.69	0.68	0.68
	SD	0.53	0.60	0.53	0.54	0.55	0.53
AlcoholCONF (31)	MAE	0.38	0.60	0.25	0.26	0.22	0.35
	MSE	-0.23	-0.48	0.09	-0.03	0.00	0.01
	MAXE	1.23	1.32	0.65	0.74	0.69	0.81
	RMSE	0.48	0.70	0.29	0.30	0.28	0.41
	SD	0.42	0.51	0.28	0.30	0.28	0.42
BeranCONF (50)	MAE	0.77	0.72	0.43	0.47	0.42	0.46
	MSE	-0.04	0.45	-0.02	-0.02	0.12	0.19
	MAXE	2.62	1.70	1.61	1.79	2.68	1.98
	RMSE	0.98	0.81	0.63	0.66	0.62	0.65
	SD	0.99	0.68	0.63	0.66	0.61	0.62
Torsion30 (2107)	MAE	0.65	0.47	0.50	0.32	0.42	0.36
	MSE	0.34	0.32	0.32	0.11	0.26	0.15
	MAXE	9.50	11.50	10.46	11.08	10.81	12.05
	RMSE	1.00	0.90	0.90	0.72	0.84	0.79
	SD	0.95	0.84	0.84	0.72	0.80	0.78
MPCONF196 (112)	MAE	3.16	1.97	2.53	1.19	3.40	1.27
	MSE	1.80	-0.64	1.64	0.50	2.60	0.90
	MAXE	14.12	7.59	12.37	4.12	15.05	4.08
	RMSE	4.44	2.47	3.61	1.53	4.75	1.59
	SD	4.07	2.40	3.23	1.45	4.00	1.32
PEPCONF-Tripeptide (647)	MAE	2.02	1.33	1.61	0.92	1.78	0.95
	MSE	1.35	-0.31	0.97	0.01	1.29	-0.14
	MAXE	9.56	5.05	7.75	3.64	8.32	4.38
	RMSE	2.61	1.67	2.09	1.17	2.30	1.19
	SD	2.23	1.64	1.85	1.17	1.91	1.19
PEPCONF-Disulfide (620)	MAE	2.52	3.52	2.42	3.05	2.60	3.12
	MSE	0.95	-2.46	0.26	-1.98	1.10	-2.11
	MAXE	10.30	20.04	13.57	20.63	9.90	20.97
	RMSE	3.18	4.69	3.16	4.28	3.29	4.34
	SD	3.04	3.99	3.15	3.80	3.10	3.79
PEPCONF-Cyclic (320)	MAE	2.85	2.31	1.81	1.77	2.30	1.53
	MSE	-0.80	-0.06	-0.09	-1.21	1.48	0.64
	MAXE	14.00	12.41	8.78	11.54	14.85	6.55
	RMSE	3.69	2.94	2.31	2.30	3.14	1.92

Data set (# of data points)		BLYP-D3/6-31G*	BLYP-D3/6-31G*-ACP	M062X/6-31G*	M062X/6-31G*-ACP	CAMB3LYP-D3/6-31G*	CAMB3LYP-D3/6-31G*-ACP
	SD	3.60	2.94	2.31	1.96	2.77	1.82
PEPCONF-Bioactive (175)	MAE	2.18	1.41	1.54	1.00	1.86	1.11
	MSE	1.39	-0.12	0.66	-0.32	1.30	-0.38
	MAXE	8.14	7.48	6.06	3.61	7.19	4.02
	RMSE	2.78	1.91	1.99	1.28	2.40	1.41
	SD	2.41	1.91	1.89	1.24	2.02	1.36
PEPCONF-Disulfide-anionic (150)	MAE	2.13	5.54	3.37	5.53	2.31	5.63
	MSE	-1.15	-4.84	-2.38	-4.65	-1.30	-4.70
	MAXE	8.44	22.17	13.67	22.47	8.38	22.48
	RMSE	2.73	7.06	4.50	7.36	2.91	7.38
	SD	2.48	5.15	3.83	5.72	2.61	5.70
PEPCONF-Bioactive-anionic (20)	MAE	1.30	0.99	1.49	1.09	1.28	1.17
	MSE	0.76	-0.01	0.77	-0.24	0.84	-0.25
	MAXE	3.94	2.40	3.36	3.75	2.67	3.37
	RMSE	1.58	1.25	1.72	1.51	1.47	1.50
	SD	1.42	1.28	1.58	1.53	1.24	1.52
ANI1ccxCONF (5254)	MAE	3.56	2.76	1.82	1.72	2.06	1.68
	MSE	-0.57	-0.29	0.76	0.61	1.09	0.90
	MAXE	29.37	20.68	17.24	12.10	22.80	11.12
	RMSE	5.04	3.71	2.97	2.28	3.64	2.25
	SD	5.01	3.70	2.87	2.20	3.47	2.06
W4-17-RE (5205)	MAE	13.00	6.52	11.13	7.67	11.17	7.32
	MSE	-0.52	-0.28	0.38	0.39	0.04	0.35
	MAXE	91.81	32.79	76.44	54.11	78.35	54.91
	RMSE	18.27	8.14	16.59	10.64	16.77	10.40
	SD	18.27	8.14	16.58	10.64	16.77	10.40
WaterOrgBH (88)	MAE	14.05	9.21	2.31	1.56	1.70	1.66
	MSE	-14.05	-9.21	0.80	-1.52	0.87	-1.36
	MAXE	18.89	12.70	5.12	4.51	5.02	4.40
	RMSE	14.46	9.34	2.63	1.82	2.12	1.90
	SD	3.46	1.56	2.52	1.00	1.95	1.33