THE DISCOVERY OF SMALL MOLECULE INHIBITORS FOR TOX1 AND ERG ONCOTARGETS WITH THE DEVELOPMENT AND USE OF PROGRESSIVE DOCKING PD2.0 APPROACH

by

Vibudh Agrawal

Btech., Indian Institute of Technology Bombay (IIT-B), 2017

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

Vancouver

October 2019

© Vibudh Agrawal, 2019
The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis/dissertation entitled:
The discovery of small molecule inhibitors for TOX1 and ERG oncotargets with the development of progressive docking PD2.0 approach.

submitted by Vibudh Agrawal in partial fulfillment of the requirements for
the degree of Master of Science
in Bioinformatics

Examinining Committee:

Dr. Artem Cherkasov
Supervisor

Dr. Michael Cox
Supervisory Committee Member

Dr. Youwen Zhou
Supervisory Committee Member

Additional Examiner

Additional Supervisory Committee Members:

Supervisory Committee Member

Supervisory Committee Member
Abstract

Drug discovery is a rigorous process that can cost up to 3 billion dollars and takes more than 10 years to bring new therapeutics from bench to bedside. While virtual screening (such as molecular docking) can significantly speed up the discovery process and improve hit rates, its speed already lags behind the rate of the explosive growth of publically available chemical databases which already exceed billions of entries. This recent surge of available chemical entities presents great opportunities for discovering novel classes of small molecule drugs but also brings a significant demand for faster docking methods. In the current thesis, we illustrated the need for a faster screening method by virtually screening 7.6 million molecules against Thymocyte selection-associated high mobility group box protein (TOX). Then we demonstrated that the deep learning-based method of ‘Progressive Docking (PD2.0)’ can speed up such virtual screening by up to hundred folds. In particular, by utilizing deep learning QSAR models trained on the docking scores of a subset of the database, one can approximate in an iterative manner the docking outcome of unprocessed entries. We tested the developed method against various targets including ETS transcription factor ERG, Estrogen Receptor Activation Function 2 (ERAF2), Androgen Receptor (AR), Estrogen Receptor (ER), Sodium-Ion Channel (Nav1.7) and many more.

In this work, we identified 18 active compounds against TOX with micro-molar potency. We also used the PD2.0 method to dock up to 1.3 billion compounds from the ZINC15 database and demonstrated that this deep-learning-based approach resulted in 65X speed acceleration and 130X Full Predicted Database Enrichment (FPDE) while retaining more than 90% of good hits. We also demonstrate the method’s robustness by docking 570 million compounds from the ZINC15 database into 13 diverse drug targets including ERG.
Lay Summary

Therapeutics discovery is an expensive process that can cost up to $3 billion dollars and more than 10 years. Nowadays, the modern drug discovery process relies on two main components: 1) computational, in which we identify potential drug candidates using virtual screening and 2) experimental, where we test these computationally identified drug candidates in actual lab settings. The objective of this thesis was to significantly accelerate the in silico identification of potential drug candidates using methods of artificial intelligence (AI). In the last 15 years, the number of synthesizable molecules has increased from less than one million to more than a billion. Thus, current computational methods of molecular docking are not fast enough to screen billions of molecules. Thus, there is a critical demand for faster VS methods and protocols and in this work, we present such an AI-enabled approach called “progressive docking (PD) 2.0”.
Preface

The following is the contribution breakdown:

i) Dr. Michael Hsing and I from Dr. Artem Cherkasov’s lab at Vancouver prostate centre are responsible for the computational part (CADD) while Dr. Mingwan Su and Yuanshen Huang from Dr. Yowen Zhou’s lab from skincare center performed experimentation for computer-aided drug discovery of small molecules against TOX (Chapter 3). Dr. Michael Hsing, Dr. Mingwan Su and I were responsible for manuscript composition and chapter 3 has been published as Agrawal V, Su M, Huang Y, Hsing M, Cherkasov A, Zhou Y. Computer-Aided Discovery of Small Molecule Inhibitors of Thymocyte Selection-Associated High Mobility Group Box Protein (TOX) as Potential Therapeutics for Cutaneous T-Cell Lymphomas. Molecules. 2019;24(19):3459.

ii) For PD2.0 (Chapter 4), the idea came from one of the previous publications of my PI Dr. Artem Cherkasov and I did most of the analysis and pipeline building (with guidance and discussions with Dr. Michael Hsing and Dr. Francesco Gentile from Dr. Artem Cherkasov’s lab at Vancouver prostate). For the part of Conformal Prediction, I got help from Dr. Ulf Norinder from Department of Computer and Systems Sciences, Stockholm University. Dr. Michael Hsing, Dr. Francesco Gentile and I were responsible for manuscript composition and chapter 4 has been published as Agrawal V, Gentile F, Hsing M, Ban F, Cherkasov A. Progressive Docking-Deep Learning Based Approach for Accelerated Virtual Screening. Paper presented at: International Conference on Artificial Neural Networks2019.

iii) I was the lead investigator for Chapter 5 and manuscripts are planned but not published.
# Table of Contents

Abstract ........................................................................................................... iii
Lay Summary ................................................................................................... iv
Preface .............................................................................................................. v
Table of Contents ............................................................................................ vi
List of Tables .................................................................................................. x
List of Figures .................................................................................................. xiii
Glossary ............................................................................................................ xx
Acknowledgments ............................................................................................ xxi

## Chapter 1: Introduction ..................................................................................... 1

1.1 Artificial intelligence a brief overview ...................................................... 1
1.2 Computer-Aided Drug Design (CADD) .................................................... 2
1.3 Molecule databases .................................................................................... 3
1.4 Mode of drug delivery ................................................................................. 4
1.5 Virtual screening (VS) ............................................................................... 5
  1.5.1 Ligand-based virtual screening .......................................................... 5
  1.5.2 Structure-based virtual screening ....................................................... 5
  1.5.3 Molecular docking ............................................................................. 6
  1.5.4 Ultra-large database docking .............................................................. 7
1.5.5 Machine learning and docking ............................................................... 7
  1.5.5.1 Scoring power ............................................................................... 8
  1.5.5.2 Ranking power .............................................................................. 9
Chapter 4: Accelerating docking using progressive docking 2.0

4.1 Introduction

4.2 Methods

4.2.1 Molecule fingerprints

4.2.2 Molecular docking

4.2.3 Clustering

4.2.3.1 K-means clustering

4.2.3.2 Mini-batch K-Means clustering

4.2.4 Logistic regression

4.2.5 Support Vector Machine

4.2.6 Random forest

4.2.7 Feedforward neural network

4.2.8 Evaluation metrics

4.2.9 Conformal prediction

4.3 Results and discussion

4.3.1 Molecular featurization

4.3.2 Database sampling

4.3.3 QSAR model for docking scores

4.3.4 Progressive docking 2.0

4.3.4.1 Training data for PD2.0

4.3.4.2 PD2.0 on Estrogen Receptor Activation Function 2 (ER-AF2)

4.3.4.2.1 Threshold-based prediction
List of Tables

Table 1. Details of protein targets selected for PD2.0. ................................................................. 18
Table 2. Top candidates for TOX small molecule inhibitors (SMIs). .............................................. 32
Table 3. Different enrichment values for MACCS, Morgan, and Pharmacophore fingerprints... 51
Table 4. Comparison of different QSAR models: Feedforward Neural Network (DNN), Random forest (RF), and logistic regression. DNN provides the best scores for all 5 enrichment values (top 10, top 100, top 1000, top 10000, FPDE at 90% recall). ................................................................. 58
Table 5. Different model statistics for threshold-based prediction after each iteration for ER-AF2. ROC-AUC, different enrichment values and recall values for the best model are reported. We observed an increase in all enrichment values after each iteration. The model cutoff/definition of good molecules improved after each iteration as well................................................................. 62
Table 6. The total number of molecules of positive class based on a fixed cutoff value in the datasets obtained at different iterations................................................................. 63
Table 7. Comparison of model statistics for threshold-based prediction between internal test set and external validation set for ER-AF2. ROC-AUC, FPDE, and recall for the best model are reported. All the statistics are consistent between the internal test set and the external validation set. ................................................................................................................................. 65
Table 8. A comparison of good molecules (TP) returned by PD2.0 and good molecules returned by docking. The predicted and docking values are very similar................................................................. 66
Table 9. Different model statistics for CP after each iteration for ER-AF2. ROC-AUC, enrichment values and recall values for the best model are reported. ................................................................. 67
Table 10. Comparing different model statistics for CP between internal test set and external validation set for ER-AF2. ROC-AUC, FPDE, and recall of the best model at each iteration are reported. ................................................................. 67

Table 11. The time required for individual steps of the PD2.0 pipeline. The major time-consuming step is docking. One PD2.0 iteration takes 3 days using 200-300 CPU cores and 4 GPUs. .................................................................................................................................................. 68

Table 12. Model ROC-AUC, FPDE, Recall for all 12 protein target for the last iteration. ........ 71

Table 13. Model cutoff, ROC-AUC, FPDE and recall for Androgen Receptor (AR) after each iteration. All the statistics are consistent between the internal and external validation set. .............. 72

Table 14. Model cutoff, ROC-AUC, FPDE and recall for Estrogen Receptor after each iteration. All the statistics are consistent between the internal and external validation set. .................. 73

Table 15. Model cutoff, ROC-AUC, FPDE and recall for Peroxisome Proliferator-Activated Receptor (PPARγ) after each iteration. All the statistics are consistent between the internal and external validation set. .............................................................................................................. 74

Table 16. Model cutoff, ROC-AUC, FPDE and recall for Calcium/calmodulin-dependent protein kinase kinase 2 (CAMKK2) after each iteration. All the statistics are consistent between the internal and external validation set. .............................................................................................................. 75

Table 17. Model cutoff, ROC-AUC, FPDE and recall for Cyclin-dependent kinase 6 (CDK6) after each iteration. All the statistics are consistent between the internal and external validation set. ........................................................................................................................................... 76

Table 18. Model cutoff, ROC-AUC, FPDE and recall for Vascular endothelial growth factor receptor 2 (VEGFR2) after each iteration. All the statistics are consistent between the internal and external validation set. ........................................................................................................................................... 77
Table 19. Model cutoff, ROC-AUC, FPDE and recall for Adenosine A2a Receptor (ADORA2A) after each iteration. All the statistics are consistent between the internal and external validation set. .................................................................................................................................................................................. 78

Table 20. Model cutoff, ROC-AUC, FPDE and recall for Thromboxane A2 Receptor (TBXA2R) after each iteration. All the statistics are consistent between the internal and external validation set. .................................................................................................................................................................................. 79

Table 21. Model cutoff, ROC-AUC, FPDE and recall for Angiotensin II type-1 receptor (AT1R) after each iteration. All the statistics are consistent between the internal and external validation set. .................................................................................................................................................................................. 80

Table 22. Model cutoff, ROC-AUC, FPDE and recall for Sodium-Ion Channel (Nav1.7) after each iteration. All the statistics are consistent between the internal and external validation set. 81

Table 23. Model cutoff, ROC-AUC, FPDE and recall for bacterial (Gloeobacter) Ligand-gated Ion Channel (GLIC) after each iteration. All the statistics are consistent between the internal and external validation set. .................................................................................................................................................................................. 82

Table 24. Model cutoff, ROC-AUC, FPDE and recall for Gamma-AminoButyric Acid receptor subunit alpha-1 (GABA α1) after each iteration. All the statistics are consistent between the internal and external validation set. .................................................................................................................................................................................. 83

Table 25. Model statistics after each iteration for ERG. ROC-AUC and FPDE values for the best model are reported. We observed an increase in all enrichment values after each iteration. The model cutoff/definition of good molecules improved after each iteration as well. ................................. 93
List of Figures

Figure 1. Two views of a molecule docked into a target pocket. (a) Protein surface representation with DNA strand (in green) and docked molecule (in pink). The red area on the protein surface are the important residues which are interacting with the molecule. (b) Same protein-molecule complex, here the blue netted surface is the defined pocket................................................................. 6

Figure 2. Computer-aided drug discovery pipeline. Different steps and tools involved in the process of computer-aided drug discovery, starting with protein structure modeling, moving on to virtual screening, candidate selection, experimental testing and finally hit-to-lead and lead optimization................................................................. 14

Figure 3. The number of synthesizable molecules available in the ZINC database over the period of the last 15 years. There is an exponential growth of the number of available synthesizable molecules in the ZINC database, from 700,000 in 2005 to more than 1 billion in 2019............. 20

Figure 4. Docking time required based on the same amount of resources for 10 million vs >1 billion molecules. Using the same amount of CPU cores (~300) to dock 10 million molecules (<1%) it takes about 7 days while it would take ~2.5 years to dock ~1 billion molecules............. 20

Figure 5. Protein structural templates for the HMG-box domain of TOX. (a) An NMR structure of mouse TOX protein (PDB ID: 2CO9) has been identified as the best structural template, with 100% sequence similarity across the 87 amino acids of the HMG-box domain, compared to the human TOX protein. (b, c) By superimposing the 2CO9 structure (orange ribbons) onto the HMG-box protein TFAM (pink ribbons, PDB ID: 3TMM, 46% sequence similarity to human TOX, 3.8A RMSD) in complex with DNA (green ribbons), the TOX-DNA interface was determined........................................................................................................................................ 30
Figure 6. Druggability of the HMG-box domain of TOX. (a) TOX-DNA interface as determined from Figure 1C. TOX HMG-box domain in orange ribbons, DNA in green ribbons. (b) Molecular surface presentation of the TOX HMG-box domain, in the same orientation as (a). (c) The TOX HMG-box domain is rotated by 180 degrees to illustrate the small-molecule binding hot spots (red). (d) A total of 200,000 drug-like molecules were docked to the TOX HMG-box domain. The percentage of interacting small molecules are shown for each protein residue as a bar graph (multiple interactions/contacts are represented as separate bars for each amino acid). Protein residues, including Gln262, Pro264, Arg273, Lys313, Glu320, Gln324 and Tyr 328 that interact with at least 10% of the small molecules, are highlighted and mapped to their corresponding locations as hot spots (red surface patches) on the TOX HMG-box domain.

Figure 7. Viability curves of TOX-high and TOX-low expressing cells. Cells were treated with various concentrations of compounds, (a) 190444, (b) 190414, (c) 190447 and (d) 190441, for 72 hours in 37°C incubator with 5% CO2. Viability was measured by CellTiter-Blue® assay and compared to the DMSO control as described in Methods and Materials. Jurkat and Hut78 cells (solid line) are the TOX-high expressing cell lines, while K562 (dotted line) is the TOX-low expressing cell line. (Produced in collaboration with Dr. Zhou’s lab)

Figure 8. Viability curves of TOX-high and TOX-low expressing cells. Cells were treated with various concentrations of compounds, (a) 190444, (b) 190414, (c) 190447 and (d) 190441, for 72 hours in 37°C incubator with 5% CO2. Viability was measured by CellTiter-Blue® assay and compared to the DMSO control as described in Methods and Materials. Jurkat and Hut78 cells (solid line) are the TOX-high expressing cell lines, while K562 (dotted line) is the TOX-low expressing cell line. (Produced in collaboration with Dr. Zhou’s lab)
Figure 9. TOX small molecule inhibitors (SMIs) bind at the hot spots on the protein-DNA interface. Docking poses are shown for compounds (a) 190444, (b) 190414, (c) 190447 and (d) 190441 on the HMG-box domain of TOX. For each panel: left) molecular surface representation, SMI in cyan, DNA in green ribbons, hot spots as red surface patches; right) in-depth molecular interactions between SMI (cyan) and protein residues of TOX (orange), hydrogen-bonds in red dotted line, hydrophobic interactions in green dotted line.

Figure 10. Different steps involved in PD2.0. Features are generated from the SMILES of the small molecule in the library and stored in a feature database. For the first iteration, a small subset of molecules (e.g. 3 million) is randomly sampled and divided into training, validation and testing set, and docked against the protein target. A score cut-off value is chosen to convert the problem from regression to classification. A QSAR model is built and used to predict the classes for all the molecules of the database (1=good and 0=bad). From the next iteration onwards, the same number of molecules (3 million) are sampled only from the positive molecules (predicted to be good binders) of the previous iteration, while ‘bad’ molecules are discarded. The validation and testing set from the first iteration remain the same for all the iterations, and the existing training set is enriched with the newly sampled molecules. All the other steps remain the same as the first iteration. This process is repeated multiple times until the total number of positively predicted molecules is a reasonable number that can be docked.

Figure 11. Docking score distribution of the entire dataset versus the docking score distribution of the largest cluster when a Tanimoto cutoff of 0.5 was used. The distributions are overlapping hence all the molecules within the cluster have the same distribution of scores as that of random sampling.
Figure 12. Accuracy and stability of the QSAR model obtained from samples of clustered and unclustered data. Each figure is a plot of FPDE (at 90% recall) against the number of clusters for different sample sizes ranging from 3000 to 480000. The mean, as well as the standard deviation, does not show any trend for any sampling technique. .......................................................... 54

Figure 13. Generalizability (criterion 1) of the QSAR model from samples of clustered and unclustered data. Each figure is a plot of the ratio between test and validation recall values against the number of clusters for different sample sizes ranging from 3000 to 480000. The mean, as well as the standard deviation, does not show any trend for any sampling technique. . 55

Figure 14. Generalizability (criterion 2) of the QSAR model from samples of clustered and unclustered data. Each figure is a plot of the ratio between test and validation precision values against the number of clusters for different sample sizes ranging from 3000 to 480000. The mean, as well as the standard deviation, does not show any trend for any sampling technique. . 56

Figure 15. The architecture of the DNN used for the QSAR model. One block is the combination of one fully connected layer and one dropout layer. The number of blocks, number of neurons and dropout frequency are the hyperparameters of the model. The input layer consists of 1024 neurons taking values from Morgan radius 2 and 1024 bits fingerprints. The output layer consists of two neurons as the problem is of binary classification.......................................................... 58

Figure 16. The plot of FPDE against training dataset size for random sampling. FPDE increases with the increasing of the dataset size.......................................................... 61

Figure 17. Generalizability (criterion 1)/recall ration for random sampling. (a) The recall ratio i.e. ratio of validation recall and test recall becomes more stable (smaller spread) with bigger sample size. (b) The standard deviation of the recall ratio goes down with a larger sample size. 61
Figure 18. Docking score distribution plot for ER-AF2 at all iterations. The score distribution shifted towards left/better molecules; the shift was more prominent at the initial iteration compared to later ones. This implies that enrichment slowed down as we proceeded through more iterations. ................................................................. 64

Figure 19. Time projection of PD2.0 and regular docking against different database sizes. As the size of the molecule database increases, the difference between the speed of PD2.0 and regular docking becomes more accentuated. Projections were generated with the assumption that when dealing with more molecules, the number of iterations, as well as inference time, will increase. Since the time per iteration is very small due to the low number of molecules required to dock (3 million), increasing by few iterations will still lead to a significant gain in speed compared to docking a library of billions of molecules. ................................................................. 69

Figure 20. Docking score distribution for AR. The distribution of scores shifted towards better (lower) docking scores after each iteration. ................................................................. 72

Figure 21. Docking score distribution for ER. The distribution of scores shifted towards better (lower) docking scores after each iteration. ................................................................. 73

Figure 22. Docking score distribution for PPARγ. The distribution of scores shifted towards better (lower) docking scores after each iteration. ................................................................. 74

Figure 23. Docking score distribution for CAMKK2. The distribution of scores shifted towards better (lower) docking scores after each iteration. ................................................................. 75

Figure 24. Docking score distribution for CDK6. The distribution of scores shifted towards better (lower) docking scores after each iteration. ................................................................. 76

Figure 25. Docking score distribution for FEGFR2. The distribution of scores shifted towards better (lower) docking scores after each iteration. ................................................................. 77
Figure 26. Docking score distribution for ADORA2A. The distribution of scores shifted towards better (lower) docking scores after each iteration. ................................................................. 78

Figure 27. Docking score distribution for TBXA2R. The distribution of scores shifted towards better (lower) docking scores after each iteration. ................................................................. 79

Figure 28. Docking score distribution for AT1R. The distribution of scores shifted towards better (lower) docking scores after each iteration. ................................................................. 80

Figure 29. Docking score distribution for Nav1.7. The distribution of scores shifted towards better (lower) docking scores after each iteration. ................................................................. 81

Figure 30. Docking score distribution for GLIC. The distribution of scores shifted towards better (lower) docking scores after each iteration. ................................................................. 82

Figure 31. Docking score distribution for GABA α1. The distribution of scores shifted towards better (lower) docking scores after each iteration. ................................................................. 83

Figure 32. Top 10, 100, 1000 enrichment for the 12 targets after the last iteration. The enrichment values decreased for an increasing number of top molecules, showing that the DNN model was able to prioritize high-affine molecules (good) i.e. molecules associated to higher DNN probabilities had higher chances of being good-scoring molecules. ............................. 84

Figure 33. The mean docking score distribution for the 12 targets across all the iterations. The mean score decreased with the iterations. The decrease rates were higher at the beginning and eventually slowed down. The rate of decrease, as well as the docking score values, were protein-dependent. ................................................................................................................................. 85

Figure 34. The number of molecules left after each iteration for the 12 targets. The number of molecules left after each iteration decreased with the iterations. The decrease was more significant at the beginning and eventually slowed down, and it was highly protein-dependent. 86
Figure 35. Top 100 enrichment value for the 12 targets after each iteration. The enrichment value increased with the iterations and the rate of increase, as well as maximum enrichment value, was protein-dependent.

Figure 36. Docking score distribution for ERG. The distribution of scores shifted towards better (lower) docking scores after each iteration. Lower docking score is an indication of a molecule with better binding affinity with ERG.
Glossary

AI Artificial Intelligence
CADD Computer-Aided Drug Design
CP Conformal Prediction
CTCL Cutaneous T Cell Lymphoma
DL Deep Learning
ER-AF2 Estrogen Receptor Activation Function 2
DNN Deep Neural Network
FPDE Full Predicted Database Enrichment
GBM Gradient Boosting Machine
LR Logistic Regression
ML Machine Learning
PCa Prostate Cancer
PD Progressive Docking
QSAR Quantitative Structure-Activity Relationship
RF Random Forest
SAR Structure-Activity Relationship
SVM Support Vector Machine
TOX Thymocyte selection-associated high mobility group box protein
VS Virtual Screening
ZINC Zinc Is Not Commercial
Acknowledgments

I like to acknowledge the following people who helped in finishing my degree:

- My supervisor Dr. Artem Cherkasov and my direct supervisor Dr. Michael Hsing for guiding me through the entire degree.
- My committee members Dr. Michael Cox and Dr. Youwen Zhou for suggesting some very important additions to my work.
- CREATE scholarship for financially supporting me through my master’s degree and Canadian Dermatology Foundation for TOX project.
- My fellow lab members Dr. Francesco Gelitle, Dr. Fuqian Ban, Satyam Bhasin, Mariia Radaeva, Godwin woo, Divya Bafna, Anh-Tien Ton and Oliver Snow for all the interesting and informative discussions.
- Special thanks to my family for always supporting me throughout my student life.
Chapter 1: Introduction

This chapter will cover all the concepts and theory required to understand and interpret the data from the upcoming chapters as well as highlight the various steps involved in the field of computer-aided drug design (CADD). It will also cover a brief introduction to artificial intelligence and how it is being integrated with the CADD pipeline.

1.1 Artificial intelligence a brief overview

Artificial Intelligence (AI) can be defined as the ability for a machine to learn, reason and correct itself just like human intelligence. Machine learning is a subset of AI and can be thought of as the process of making a computational algorithm to learn the embedded structure from data. Data can be images, prices, costs of goods, gene expression data or any data which may have hidden patterns. Deep learning (DL), in particular, being a sub-discipline of machine learning, works on extracting higher-level information from raw data in multiple steps/layers. The idea is that each subsequent step/layer learns a more abstract and composite representation of the raw data, compared to the previous layer. The other way to look at DL is to consider various types of supervision applied to the deep learning phase, which can be divided into three categories: fully-supervised/supervised, semi-supervised and unsupervised. Fully supervised and unsupervised, as the name suggests, require full supervision/labeled data and no supervision/unlabeled data respectively. Some of the examples of supervised machine learning models include (but are not limited to) Linear Regression\(^3\), Logistic Regression\(^4\), Random Forest\(^5\) (RF), Support Vector Machines\(^6\) (SVM), Gradient Boosting Machines\(^7\) (GBM), Neural Networks\(^8\) (NN) methods. Methods of unsupervised learning encompass clustering\(^9,10\) and dimensionality reduction\(^11,12\) among others.
Examples of supervised deep learning models that gained particular attention in recent years include Deep Feed Forward Neural Network\textsuperscript{13} (DFFNN), Convolution Neural Network\textsuperscript{14} (CNN), Recurrent Neural Network\textsuperscript{15-17} (RNN), while popular Variation Auto Encoders\textsuperscript{18} (VAE) and Generative Adversarial Network\textsuperscript{19} (GAN) correspond to unsupervised DL.

The semi-supervised learning, on the other hand, doesn’t require the exact labels but requires a criterion to reward the machine if it does the correct task and lets it figure out how to do it correctly based on these rewards (very similar to humans). AI in general is the combination of all three, supervised, unsupervised and semi-supervised learnings. Reinforcement learning is a semi-supervised method and is part of AI, in which the machine learns, reason and correct itself based on the things it observes, some of the examples are machine learning to play games, robots able to learn to walk and function, a self-driving car able to make decisions. For all of these tasks, just like humans, machines learn from experience as it makes mistakes.

Out of the three types of learning, supervised one is the most broadly used and studied. Some of the fields where supervised learning made a significant impact in recent years include object detection\textsuperscript{20, 21}, machine translation\textsuperscript{22, 23}, speech recognition\textsuperscript{24}, facial recognition\textsuperscript{25, 26}, biomedical imaging\textsuperscript{27}, as well as predictive pharmacology and computer-aided drug discovery (CADD) among others\textsuperscript{28}.

1.2 Computer-Aided Drug Design (CADD)

Drug discovery is an expensive and very slow process that can cost up to $3 billion dollars and takes more than 10 years to bring a drug from laboratories to clinical practice. Two of the main challenges are 1) low hit discovery rate ($< 1\%$), when conventional chemical libraries are screened \textit{in vitro} against a disease-associated target, and 2) unwanted side-effects arisen from hit
compounds cross-reacting with other protein targets\textsuperscript{29,30}. The CADD can significantly speed up compounds screening and improve hit rates by imposing virtual screening (VS) protocols to millions of drug-like small molecules. VS scores those molecules against a specific site on a target protein \textit{in silico} and triages them for subsequent experimental validations\textsuperscript{30}. Such VS protocols allow achieving >10\% hit (success) rates in initial drug discovery campaigns, as opposed to \textasciitilde0.03\% yields of wet-lab screening\textsuperscript{29}.

The other core steps in computer-aided drug design campaign\textsuperscript{30} include compound design, energy calculations, SAR analysis, ADME modeling, and drug-target interaction prediction\textsuperscript{31}. Below we present the high-level overview of all those steps.

1.3 Molecule databases

Current molecular databases contain diverse information about known (or sometimes, prospective) compounds, their activities (expressed as IC50 numbers, binding affinities, etc), physicochemical properties, relevant information on chemical vendors, etc. Some of the most well-known databases are ZINC1\textsuperscript{52} (consists of purchasing information for more than a billion molecules), ENAMINE\textsuperscript{33} (consists of purchasing information for 100s of million molecules), PubChem\textsuperscript{34} (consists of experimental information on millions of molecules-target pairs), PDB Database\textsuperscript{35} (stores 3D coordinates of proteins, nucleic acid and complex assemblies), BindingDB\textsuperscript{36} and PDBbind\textsuperscript{37} (cataloguing binding affinity number of millions of ligand-target pairs and relevant information on the corresponding experimental conditions, etc).
1.4 Mode of drug delivery

The drug can be introduced into the body in several ways including orally, through the skin (transcutaneous), through injections directly to the bloodstream, placed under the tongue, inserted in rectum or vagina, placed in the eye, etc. In this thesis, we focus on the oral and transcutaneous mode of delivery.

The oral route is the most convenient, safest as well as inexpensive for delivering the drug. The drug can be in the form of tablets, capsules or liquids. The absorption of drugs happens in the mouth, stomach and small intestine. Out of the three most drugs get absorbed in the small intestine. Intestinal wall alters many drugs chemically, causing only a small amount of drug to be absorbed and reach the bloodstream, thus oral delivery requires a high dosage of the drug. There are several properties that we expect in an oral drug based on the existing drug candidates including lower molecular weight (<500 Daltons), lower rotatable bond count, lower hydrogen bond counts, lower surface area and certain minimum lipophilicity (octanol-water partition coefficient)\textsuperscript{38}.

The transcutaneous route of delivery is the delivery of drugs through the skin, drugs can be delivered through a patch on the skin. The drug needs to penetrate the skin to reach the bloodstream, for enhancing the skin penetration these drugs are sometimes mixed with other chemicals such as alcohol. The patches can stay for hours or days delivering the drug slowly but continuously. This form of drug delivery is useful for drugs that are removed from the system very fast thus needed to be taken frequently. Patches, on the other hand, can cause skin irritation and can only be used for drugs that need to be given in small daily doses. Some desirable properties for such drug candidate include lower molecular weight, lipophilicity (octanol-water partition coefficient) greater than 1, lower number of hydrogen bonds since permeability across lipid bilayer decreases with more hydrogen bonds, lower number of hydrogen acceptors since too many
hydrogen bond acceptors hinders the permeability across the lipid bilayer and molecules with no charge\(^{39-41}\).

1.5 Virtual screening (VS)

VS is an efficient approach for the identification of hit compounds for a given protein target. The starting point for the VS campaign is the identification of a source of already available or readily synthesizable chemicals, such as ZINC database\(^{32}\), which consists of billions of molecular entries. The next step is to apply VS protocol(s) such as ligand-based or structure-based screening, depending on the target structure is available or reliably modelable.

1.5.1 Ligand-based virtual screening

For ligand-based screening, a similarity search is performed between known active molecules and the entire small molecule database library. The search is based on some form of representation of small molecules also known as fingerprints or descriptors. One of the drawbacks of such an approach is that the calculated similarity relationship does not have any strong correlation with the observed activity and hence results in a high number of false positives.

1.5.2 Structure-based virtual screening

Structure-based virtual screening is arguably the most accurate VS protocol, but its use is only feasible when a target protein possesses an experimentally resolved structure. This structure or closely related homology models are then used to identify a targetable pocket or a ligand-binding site to be inhibited. Once the target pocket is defined (some examples are shown in Figure
1), by a certain CADD method, such as chemical probing or hot-mapping, a significant number of small molecules can be screened against this pocket using means of molecular docking.

![Figure 1](image)

**Figure 1.** Two views of a molecule docked into a target pocket. (a) Protein surface representation with DNA strand (in green) and docked molecule (in pink). The red area on the protein surface are the important residues which are interacting with the molecule. (b) Same protein-molecule complex, here the blue netted surface is the defined pocket.

1.5.3 Molecular docking

The process of molecular docking involves the search for the most energetically favorable binding pose of a given compound within a defined pocket. In a nutshell, the docking workflow can be divided into two parts 1) performing a conformal search across all possible conformers and 2) calculating the binding free energy of each conformer using a VS scoring function. The conformal search problem is a solved case for docking protocols, but it is difficult to accurately rank compounds based on predicted binding free energy (scoring function). Of note, the current
scoring functions which only approximate the binding free energy for a protein-ligand pair represent the limiting step in docking and have significant room for improvement.

1.5.4 Ultra-large database docking

In one of the recent papers by Lyu et al\textsuperscript{42}, authors docked a library of \(\sim\)100 million molecules into two targets - AmpC\(\beta\)-lactamase (AmpC) and D\textsubscript{4} dopamine receptor. They discovered that for these targets, their known ligands (active molecules) appear among the top hits i.e. the used docking methods could correctly prioritize a very large chemical space. They compared the docking of a small library of compounds with the docking of a larger library and observed that performance is typically improved in the latter case. Authors concluded that the ultra-large libraries contain molecules that are better suited for a given receptor structure and that by searching large libraries one can discover better drug candidates.

Furthermore, authors anticipated that given an ultra-large library, it would be logical and time-saving to cluster them and only dock particular cluster representative(s).

1.5.5 Machine learning and docking

As outlined above, the process of docking can be divided into two major steps: 1) conformal search and 2) free energy estimation. The latter phase is particularly error-prone and computationally extensive. The free energy estimation in a docking process is typically approximated by a much-simplified scoring function. The good scoring function is expected to have three properties: 1) scoring power, i.e. the ability to correctly predict the binding affinity for different binding poses 2) ranking power or ability to rank the binding poses from a set of ligands with known poses against the same target and 3) docking power - the ability to identify the best
binding pose of a given ligand amongst different generated conformers\textsuperscript{43}. In the next section, different statistical and machine learning methods are discussed in the context of the above mentioned three properties.

1.5.5.1 Scoring power

Scoring power is defined as the ability of a scoring function to predict binding energy for a protein-ligand complex with known 3D coordinates. Ideally, the corresponding predicted scores should be linearly correlated with experimentally measured binding affinities. There are three categories of scoring functions\textsuperscript{44}: 1) force field-based 2) empirical and 3) knowledge-based. All three are parametric methods that involve estimating the fixed number of parameters based on the available experimental data. Thus, the force field- methods estimate the parameters of molecular mechanics energies approximated from experimental data and/or ab initio simulations\textsuperscript{45}. For empirical scoring functions, the complete energy term is composed of individual weighted energy terms, where the weight coefficients obtained from a regression on experimental binding energies\textsuperscript{46}. Lastly, knowledge-based functions are the weakest predictors and based on the theory that a large database of protein-ligand complexes can be statistically mined to deduce rules and models implicitly embedded in data\textsuperscript{47}.

Various machine learning methods are actively used for constructing VS scoring functions. Thus, Kinnings et al.\textsuperscript{48} used Support Vector Machine (SVM) to derive individual weight terms of different protein families for the empirical scoring function. The same method can be used for deriving different parameter coefficients for force-field based scoring function. There are some ML approaches that used protein-ligand features available in the literature (geometric features, pharmacophore features, physical force field energy terms) for predicting the binding affinities.
Few other non-parametric models tried to learn non-linear dependency of the structure of the protein-ligand complex with the binding affinity\textsuperscript{49}. In this approximation, each feature represents the number of occurrences of a particular protein-ligand atom type pair interacting within a certain distance range. The authors used the random forest to model the relationship between these features and binding affinity. Overall the features can be divided into 1) physical features (e.g. energy terms) 2) chemical features (fingerprints) 3) Geometric features (Based on the structure of protein-ligand complex).

The work by Li et.al\textsuperscript{50} used RF with a combination of different energy as well as geometric features and showed that incorporating more features and training with more data can boost the performance of the model. In the work by Ballester et.al\textsuperscript{51}, authors utilized chemical descriptors to represent complex and showed that having more precise chemical descriptors does not always lead to a more accurate model. Recently, there had been a number of similar publications on scoring functions and machine learning\textsuperscript{43, 52}.

1.5.5.2 Ranking power

Ranking power is defined as the ability of a scoring function to properly rank ligands for their binding affinity to the same protein. There are two classical approaches for evaluating the ranking power: 1) high-level ranking and 2) low-level ranking. The PDBbind\textsuperscript{37} v2013 database contains complexes with 3 different ligands against the same protein target. The high-level ranking is defined by correctly ordering the three ligands based on the binding affinity towards the protein target. The low-level ranking on the other hand just requires to correctly identify the best binding ligand out of the three. For any protein target, one point is awarded if the scoring function succeeds for high-level ranking/low-level ranking.
In another study, authors evaluated a panel of 20 conventional scoring functions and demonstrated that the methods called the “S-score” ranked the ligands with the highest accuracy in both high-level and low-level ranking exercises. The work presented by Ashtawy et al. assessed the ranking power of ML-based function on PDBbind 2007 and 2010 benchmark datasets. The authors also used a very diverse set of molecule features: X-score, AffiScore, and RF-score. For ML models they used Multivariate Linear Regression (MLR), multivariate adaptive regression splines (MARS), k-nearest neighbours (KNN), SVM, RF, and Boosting Regression Trees (BRT). Out of all these, they found ensemble-based methods like RF and BRT worked the best, with RF having the best results for high-level ranking (62.5%) as well as low-level ranking (78.1%). Some of the other works involved using non-parametric ML models such as inductive logic programming (ILP) combined with SVM; and SVR-based scoring functions: 1) SVR-knowledge based and 2) SVR-empirical descriptor based. More work can be found in the review papers on machine learning and docking.

1.5.5.3 Docking power

The docking power is defined as the ability of a scoring function to identify the native binding pose amongst the generated ones. In other words, it should score the native pose as the best among all possibilities. The classical way of testing a scoring function is to generate some decoy binding poses (usually few 100s) and include the native binding poses in them to see it gets scored on the top and whether the RMSD between the poses is low (<2.0Å or so). Out of 20 scoring functions tested by the authors of a review, paper ChemPLP@GOLD and Chemscore@GOLD resulted in a success rate above 80%.
Finding the right conformation is usually done using: steepest descent optimization, Monte Carlo Simulation, simulated annealing, Molecular Dynamics (MD), genetic algorithms, and geometry methods. The importance of docking power of any scoring function has been demonstrated previously\(^6\); there is a variety of ML-based scoring functions that are in-sensitive to docking pose accuracy. Thus, authors of \(^6\) compared the docking power of linear and ML-based scoring functions and demonstrated that the docking power of linear scoring functions is significantly higher than those of ML-based scoring functions. In other study\(^6\) researchers had used deep learning approximation to predict the binding affinity values by extracting features from the poses obtained from two docking programs.

### 1.5.6 Predicting docking score using machine learning

We have briefly reviewed a variety of methods published to date on the prediction of protein-ligand binding affinities with various scoring functions. Yet, Progressive Docking (PD)\(^6\) was the first reported method that utilized classical QSAR methodology to simulate docking scores and further use them for reducing the number of remaining docking jobs in a CADD pipeline. The original idea of PD was to build model specific to the protein target site using a subset of docked molecules and to predict the scores for all the remaining molecules using QSAR model with 3D descriptors. The predicted docking scores are then used to iteratively removing less promising molecules from undocked database. In the last few years there have been few similar studies aimed to approximate docking scores using methods like SVM, RF enhanced with conformal prediction statistics\(^6,\)\(^6\). These methods used the same idea of PD i.e of progressive removal of undocked molecules based on some classification (instead of quantitative regression, as in case of PD).
1.5.7 **Structure-Activity Relationship (SAR) and lead optimization**

As the name suggests SAR methods link a structure of protein-ligand complex to some measure of ligand’s activity like the corresponding IC50 value. Historically, the field of Quantitative Structure-Activity Relationship (QSAR) was encompassing ligand-based approaches which are important components of computer-aided drug discovery. The origin of QSAR dates back to the 1960s and was developed by Corwin Hansch. QSAR model can be linear or non-linear in nature, the advantage of the former class is in their interpretability, while they are generally less predictive.

There are also various SAR visualization techniques which are more intuitive and attempt to compliment analog design with the graphical representation of SAR patterns and identify key components for choosing new analogs. Nowadays, QSAR models play a very important role in hit-to-lead and lead optimization workflows.

In such workflows, after the initial active compounds (hits) is identified, the best ones undergo hit-to-lead optimization to qualify as the lead candidates, which is further optimized into drug candidates. Among other steps, the hit-to-lead optimization involves the search for hit analogs in chemical space in an iterative manner and using QSAR approaches.

After the lead compound has been selected based on the hit-to-lead process the lead compound move to the lead optimization process. In this process, the lead compound is iteratively made more potent and have less off-target effects by medicinal chemistry knowledge. New custom compounds are made based on the lead molecule and their activity is predicted using the QSAR model and eventually, these molecules are purchased and tested in vitro.
1.5.8 **ADMET Prediction by QSAR**

ADMET stands for “absorption, distribution, metabolism, excretion, and toxicity” and describes the main pharmaceutically-relevant characteristics of a compound administered into a living organism. ADMET modeling usually involves building a variety of machine learning models as well as statistical modeling and is also one of the very hot topics in the field of computer-aided drug discovery. Among other applications, ADMET involves predicting the complex in vivo behavior of bioactive compounds which we do not understand completely ourselves. The data available for ADMET prediction is very sparse making it even more difficult for building good models. Over the last few years, there was a variety of computational methods developed for elaborate prediction of drug metabolism and toxicities\(^{76,77}\).

1.5.9 **In-house CADD pipeline**

We had developed our own computer-aided drug design pipeline\(^30\) (Figure 2) which had previously led us to successfully discover SMI’s for other cancer drug targets such as AR\(^78\), ERG\(^79\), and MYC\(^80\) among others\(^81-83\). Most of the steps are same as every other CADD pipeline but also involve a few additions to some of the intermediate steps.
Figure 2. Computer-aided drug discovery pipeline. Different steps and tools involved in the process of computer-aided drug discovery, starting with protein structure modeling, moving on to virtual screening, candidate selection, experimental testing and finally hit-to-lead and lead optimization.

The first step is protein structure modeling which is part of every CADD pipeline, in where we get the crystal structure of the protein from the PDB database (if the structure is available) or performs homology modeling on PDB-retrieved templates. We then use MOE\textsuperscript{84} software for identifying the best targetable pocket for the protein which can be used in the subsequent step of virtual screening.

At the next stage, we perform virtual screening of the ZINC database. Previously we used to work with the entire ZINC as it contained 10-20M entries. But these days ZINC database has grown more than 1 billion entries and, it would take too long to screen all of them. Thus, we apply various filters to ZINC such as numbers of Hydrogen bonds, Rings, LogP etc\textsuperscript{85} to decrease the dataset to 10’s of million manageable molecules.

For molecular docking, we use GLIDE\textsuperscript{86}, OEDDocking\textsuperscript{87}, and ICM\textsuperscript{88} packages. After the docking round, we select molecules that have consistent pose across all the three software based on the corresponding threshold RMSD value of 2A. We also calculate other properties of virtual
hits such as ADMET\textsuperscript{77} characteristics, drug likeness\textsuperscript{89, 90} and predicted pK\textsubscript{i}. Based on all these calculated properties we derive a consensus score and select top 100’s of molecules and do clustering to remove similar structures.

Finally, we select ~100 molecules for testing experimentally (in vitro). Based on the hit compounds we do hit-to-lead and lead optimization using QSAR models, 2D/3D similarity search and Molecular Dynamic (MD) simulations.

1.6 Protein targets

1.6.1 Thymocyte selection-associated high mobility group box protein (TOX)

Thymocyte selection-associated high mobility group box protein (TOX) is a 526 aa nuclear protein (molecular weight, 57 kD) that binds DNA in structure-dependent and sequence-independent manner\textsuperscript{91}. It is a member of the evolutionarily conserved high mobility group (HMG) box family and a key regulatory nuclear protein in the development of CD4\textsuperscript{+} T cells, natural killer cells, and lymphoid tissue inducer cells\textsuperscript{91-95}. The roles of TOX in immune system development are well-characterized\textsuperscript{91}. TOX expression is tightly controlled and it is expressed during early CD4\textsuperscript{+} T cell development. The importance of TOX in T cell development and maturation has been shown before by gene knockout experiments of TOX on mice. In recent years, strong evidence has emerged that TOX is a specific biomarker, strong prognostic factor, key pathogenic driver, and attractive therapeutic target for CTCL\textsuperscript{96-99}.

1.6.2 ETS transcription factor ERG

ERG (ETS-related gene) is a transcription factor with a full-length of 486 amino-acids and a molecular weight of 54kDa\textsuperscript{100}. ERG is a member of the ETS family, the majority of which have
an ETS DNA-binding domain (DBD) that is about 85 amino acid long. This DBD recognizes the DNA sequence containing a core GGA(A/T) motif. ERG is initially highly expressed in the embryonic mesoderm and endothelium where it plays an important role in bone development, the formation of urogenital tract and vascular system. ERG is also expressed at high levels in embryonic neural crest cells during their migratory phase. ERG expression decreases during vascular development but continues to regulate the pluripotency of hematopoietic stem cells, endothelial cell (EC) homeostasis and angiogenesis.

Genomic alteration involving translocation of the ERG occurs in approximately half of the prostate cancer cells. These alterations result in aberrant, androgen-regulated production of ERG protein variants that directly contribute to disease development and progression. ERG modulates various PCa related phenotypes such as disruption of the epithelial differentiation program via AR dysregulation, activation of c-Myc, epigenetic reprogramming via EZH2 and promotion of genomic instability via PARP dysregulation. Overexpression of ERG also promotes epithelial-mesenchymal transition (EMT) and enables the transformed cells to acquire migratory and invasive characteristics. Previously, a new class of anti-ERG small molecule inhibitors (compound VPC-18005 and its derivatives) have been developed at Vancouver Prostate Centre through the use of an in-house CADD pipeline. In summary, there are multiple lines of evidence that together strongly demonstrate that ERG is an attractive molecular target for developing PCa therapies.

1.6.3 Other targets

To evaluate the performance of the developed PD2.0 approach we considered 12 proteins from four main drug-targeted families: nuclear receptors, kinases, G-coupled protein receptors,
and ion channels\textsuperscript{113}. These targets are very well-validated and also have well-defined x-ray structures (where the target is bound to various ligands), and most of them, already have approved drugs. Details of these proteins are presented in Table 1.

We also considered Estrogen Receptor Activation Function 2 (ERAF2) as a target which is an unconventional pocket of Estrogen Receptor (ER), which undergoes the \textit{in house} drug development. This hydrophobic cavity on the surface of ER\textalpha is involved in the recruitment of different coactivator proteins of ER\textalpha, and a number of recent studies have successfully explored ER-AF2 as an alternative target site for ER inhibitors to the hormone-binding pocket in breast cancer.
Table 1. Details of protein targets selected for PD2.0.

<table>
<thead>
<tr>
<th>Family</th>
<th>Target</th>
<th>Ligand</th>
<th>PDB ID</th>
<th>Resolution</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclear receptors</td>
<td>AR</td>
<td>Dihydrotestosterone</td>
<td>1T7R</td>
<td>1.40 Å</td>
<td>Prostate cancer</td>
</tr>
<tr>
<td></td>
<td>ERα</td>
<td>Raloxifene</td>
<td>1ERR</td>
<td>2.60 Å</td>
<td>Breast cancer, osteoporosis, menopausal symptoms</td>
</tr>
<tr>
<td></td>
<td>PPARγ</td>
<td>Rosiglitazone</td>
<td>5YCP</td>
<td>2.00 Å</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Kinases</td>
<td>CAMKK2</td>
<td>STO-609</td>
<td>2ZV2</td>
<td>2.40 Å</td>
<td>Prostate cancer, metabolic hepatic diseases</td>
</tr>
<tr>
<td></td>
<td>CDK6</td>
<td>Abemaciclib</td>
<td>5L2S</td>
<td>2.27 Å</td>
<td>Breast cancer</td>
</tr>
<tr>
<td></td>
<td>VEGFR2</td>
<td>Axitinib</td>
<td>4AG8</td>
<td>1.95 Å</td>
<td>Multiple cancer types</td>
</tr>
<tr>
<td>G-protein coupled</td>
<td>ADORA2A</td>
<td>Theophylline</td>
<td>5MZJ</td>
<td>2.00 Å</td>
<td>Myocardial perfusion imaging, inflammation, neuropathic pain, Parkinson’s</td>
</tr>
<tr>
<td>receptors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>disease</td>
</tr>
<tr>
<td></td>
<td>TXA2</td>
<td>Ramatroban</td>
<td>6IU</td>
<td>2.50 Å</td>
<td>Cardiovascular diseases, asthma</td>
</tr>
<tr>
<td></td>
<td>AT1R</td>
<td>Olmesartan</td>
<td>4ZUD</td>
<td>2.80 Å</td>
<td>Hypertension</td>
</tr>
<tr>
<td>Ion channels</td>
<td>Nav1.7</td>
<td>GX-936</td>
<td>5EK0</td>
<td>3.53 Å</td>
<td>Pain</td>
</tr>
<tr>
<td></td>
<td>GLIC</td>
<td>Anesthetic ketamine</td>
<td>4F8H</td>
<td>2.99 Å</td>
<td>General anesthetics</td>
</tr>
<tr>
<td></td>
<td>GABA α1</td>
<td>THDOC</td>
<td>5OSB</td>
<td>3.80 Å</td>
<td>Neurological disorders</td>
</tr>
</tbody>
</table>
Chapter 2: Problem and specific aims

2.1 Problem

Purchasable chemical databases such as ZINC have grown exponentially from 700,000 compounds in 2005\textsuperscript{126} to over 1 billion molecules in 2019\textsuperscript{32} (Figure 3). Previously due to limitations in available computer power and methods, we could only virtually screen around 10 million molecules (representing less than 1% of the entire ZINC database as shown in Figure 4) through the use of molecular docking. As it has been shown before\textsuperscript{42}, the larger the chemical libraries screened, the better the chance of finding optimal drug candidates. Thus, these billions of molecules in chemical databases such as ZINC provide a tremendous opportunity for drug discovery, provided that if we could accelerate and scale up molecular docking methods from tens of millions to billions of compounds. We encountered such a challenge from drug discovery projects targeting the TOX and ERG proteins.
Figure 3. The number of synthesizable molecules available in the ZINC database over the period of the last 15 years. There is an exponential growth of the number of available synthesizable molecules in the ZINC database, from 700,000 in 2005 to more than 1 billion in 2019.

Figure 4. Docking time required based on the same amount of resources for 10 million vs >1 billion molecules. Using the same amount of CPU cores (~300) to dock 10 million molecules (<1%) it takes about 7 days while it would take ~2.5 years to dock ~1 billion molecules.
2.2 Aims and objectives

In this thesis, we describe the development of a new virtual screening pipeline ‘Progressive Docking PD2.0’ based on a deep learning framework, which can accelerate the docking process for very large chemical databases (>1 billion molecules). The PD2.0 uses docking scores generated from a small subset of the molecules, to predict binding affinities of the remaining molecules in the entire chemical library via a progressive, iterative manner.

1. **Identify novel small molecule inhibitors targeting TOX through the use of virtual screening.** There are currently no small molecules inhibitors of TOX, and herein we aimed to address this unmet need by developing anti-TOX therapeutics through our in-house CADD pipeline. (While this aim only utilizes the docking of 7 million molecules, it motivates us to fully develop PD2.0)

2. **Develop the Progressive Docking PD2.0 method to validate its performance on a diverse list of drug targets.** For molecule features, we evaluated MACCS, Morgan, and Pharmacophore fingerprints. For choosing the QSAR model we compared Deep Neural Networks (DNN) with other machine learning models, including Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM). Then we selected the sampling technique based on performance and time complexity. We then optimized the pipeline by testing various ways to combine different components within the pipeline. Lastly, we validated the PD2.0 on ERAF2 and 12 targets mentioned in Table 1.

3. **Apply PD2.0 on ERG for the discovery of new small molecule inhibitors.** Previously, we have applied CADD to virtually screen 3 million molecules from the ZINC database and developed a list of anti-ERG inhibitors\(^7^9\). We have now virtually screened 350 million
molecules with the use of the PD2.0 method on ERG to identify new classes of small molecule inhibitors.
Chapter 3: Computer-aided discovery of small molecule inhibitor of thymocyte selection-associated high mobility group box protein (TOX)

3.1 Introduction

In this chapter, we are going to demonstrate the need for a faster computational method for virtual screenings. In particular, we identified a prospective protein target TOX for Cutaneous T cell lymphoma (CTCL) and found new SMI’s to target this protein using virtual screening followed by experimental validation. CTCL is a primary lymphoma of the skin and is derived from cutaneous resident memory T cells. In most of the cases, malignant T cells are CD4+ i.e. CD4 (cluster of differentiation 4), which is a glycoprotein is expressed in those cells. The most common variants of CTCL are mycosis fungoides (MF) and Sezary syndrome (SS). Only about 10% of CTCL patients with minor conditions and about 25% with major conditions end up developing end-stage conditions such as leukemic stage CTCL including SS, but for the rest, the life span is that of a healthy individual. SS patients have a median survival of 2-4 years and an estimated 5-year survival rate of 24%\textsuperscript{127}.

Current treatment of CTCL depends upon the level of severity and clinical stage, for example, the treatments available for early stages are skin-directed, including topical steroids, topical nitrogen mustard, topical retinoid as well as phototherapy. It gets more difficult to treat late stages when the disease is no longer limited to the skin. Although a number of agents, including retinoids, interferon (IFN), monoclonal antibodies, epigenetic modifiers such as histone deacetylase inhibitors (HDACi), and denileukin diftitox, have shown benefit in the treatment of advanced disease, none are curative. Therefore, new therapies for CTCL are urgently needed.
Thymocyte selection-associated high mobility group box protein (TOX) is a 526 aa nuclear protein (molecular weight, 57 kD) that binds DNA in a structure-dependent and sequence-independent manner\textsuperscript{91}. It is a member of the evolutionarily conserved high mobility group (HMG) box family and a key regulatory nuclear protein in the development of CD4+ T cells, natural killer cells, and lymphoid tissue inducer cells\textsuperscript{91-95}. The roles of TOX in immune system development are well-characterized\textsuperscript{91}. TOX expression is tightly controlled and it is expressed during early CD4+ T cell development. The importance of TOX in T cell development and maturation has been shown before by gene knockout experiments of TOX on mice. In recent years, strong evidence has emerged that TOX is a specific biomarker, strong prognostic factor, key pathogenic driver, and attractive therapeutic target for CTCL\textsuperscript{96-99}. In particular:

(1) TOX is aberrantly expressed in CTCL: In comparative transcriptome studies\textsuperscript{99}, TOX emerged as the most highly enriched in CTCL skin biopsies.

(2) Enhanced transcript levels of TOX correlate with increased risk of disease-specific mortality in CTCL: Further experiments showed that TOX expression levels in CTCL skin biopsies and in peripheral blood purified malignant CTCL cells were positively correlated with disease-specific mortality of CTCL patients\textsuperscript{98}.

(3) Stable knockdown of TOX inhibits the growth of CTCL cells in vitro: TOX gene silencing suppresses CTCL cell proliferation as well as decreases colony formation capability in vitro\textsuperscript{97}.

(4) TOX suppression induces apoptosis and caspase activation in CTCL cells: With TOX suppressed, CTCL cells had increased apoptotic cells\textsuperscript{97}.
(5) TOX suppression impairs the tumor-forming ability of CTCL cells in vivo: Injecting CTCL patient-derived cell into Mouse readily forms a tumor but this is not observed when TOX expression was silenced\textsuperscript{97}.

(6) TOX suppression led to expression changes of multiple downstream genes: TOX silencing led to a change in expression levels of numerous genes out of which SMAD3 is the most significant. SMAD3 is usually suppressed in the presence of TOX but sensitively induced after TOX gene silencing\textsuperscript{97}. In summary, there are multiple lines of evidence that together strongly demonstrate that TOX is an attractive molecular target for developing CTCL therapies.

There are currently no small molecules inhibitors of TOX, and herein we aimed to address this unmet need by developing anti-TOX therapeutics through a computer-aided drug design (CADD) platform that we have previously established\textsuperscript{30} and successfully utilized in a number of other cancer-related drug targets including AR\textsuperscript{78}, ER\textsuperscript{83}, ERG\textsuperscript{79}, MYC\textsuperscript{80}, etc. Here we report the use of this established CADD pipeline, which combined virtual screening of 7.6 million drug-like small molecules within vitro experimental validation, to discover new classes of anti-TOX compounds.

3.2 Methods

3.2.1 Structural evaluation of TOX druggability

Since there is no resolved human TOX NMR structure available, we found mouse TOX (PDB ID: 2CO9) as the best matching template (100% sequence similarity) for human TOX (from K251 to Y337, with residue numbering based on UniProt ID: O94900-1). We then identified the suitable binding site on human TOX model by docking 200,000 small molecules (ZINC database\textsuperscript{32}, in-stock, 3D representation, with molecule weight from 375 to 400 Dalton, and logP = 1.5 +/- 1)
against the HMG-box domain. We used the “blind docking” setup of GLIDE\textsuperscript{86} i.e. no binding site was defined ((Standard Precision mode with default parameters) from Schrodinger 2016-3 package). For identifying the important residues i.e. the residues with the frequency of interaction with the molecules of greater than or equal to 10\%, we applied Protein-Ligand Interaction Fingerprints (PLIF) on the top hit compounds using Molecular Operating Environment (MOE)\textsuperscript{84}.

3.2.2 In silico screening

For initial virtual screening of probing 200,000 molecules, we took the top 10\% scored entries (20,000 molecules) predicted by GLIDE and applied the in-house pipeline described in Chapter 1. In particular, these 20,000 molecules were docked using eHiTS program using the “blind docking” setup and the molecules with consistent pose between GLIDE and eHiTS (RMSD\(<=3\AA, 136 \)molecules) were docked using ICM (default parameters). Among the processed molecules, 22 demonstrated docking consistency with all three programs (Glide, eHiTS, and ICM). Through in vitro testing on TOX expressing Sezary cell lines, one of the candidates, VPC-190010, showed inhibition activity with IC50 less than 50 \(\mu\)M in all the TOX expressing Sezary cell lines (data not shown).

For the large-scale docking (7.6 million molecules), the binding site was identified using the docking pose of initial hit compound VPC-190010 (TOX hot spot residues including Gln262, Pro264, Arg273, Lys313, Glu320, and Gln324). The screening process was divided into 1) compounds for the oral mode of delivery and 2) topical model of delivery. All the molecules with 3D representation and logP\(>-1\) were downloaded from the ZINC database.
For the oral application, the filters specific to the oral mode of delivery such as molecular weight> = 350 Dalton filter specific to drug-like criteria from FA-Drugs4 and filter based on the initial in-silico screening such as number of rings between 4 and 6 were applied. As a result, a total of ~ 3 million molecules were retained and then docked with Glide (Standard Precision mode with default parameters). Molecules with docking scores less than -5 (lower is better) were further docked with the FRED program, OEDocking (up to 500 conformers were generated for each molecule and were docked using FRED with default parameters), and the corresponding RMSD values were calculated for the top poses. All the molecules with an RMSD ≤ 3Å were retained and were docked again using ICM. For the poses predicted by ICM (default parameters), RMSD values against Glide were calculated and only the molecules with RMSD ≤ 3Å were retained.

Within the resulting set, theoretical pKi values were calculated for each molecule using a custom MOE SVL script. Other properties like ADMET (absorption, distribution, metabolism, excretion, toxicity) and pharmacokinetics predictions were also calculated by using computational programs such as ADMET Predictor, FAF-Drugs and Quantitative Estimate of Drug-likeness (QED).

In the next step, a consensus scoring method was used and Molecules with total vote greater or equal to 7 were retained and then clustered together to remove the similar compounds (70% similarity). Finally, a total of 66 compounds were chosen for experimental validation.

For the topical application candidate compounds, the filters specific to the topical mode of delivery were applied: (MW<350 Dalton, charge = 0 and 2 ≤ LogP ≤ 4) along with filters specific to lead-like criteria such as FA-Drugs4. Apart from the above-mentioned criteria, molecules with chiral centers ≤ 1 (vendors usually sell the racemic mixture), 2 ≤ number of rings ≤ 4 (molecules with 1 ring are too simplistic), rotatable bonds ≤ 6, and 2 ≤ hydrogen bond acceptors ≤ 7 were
retained. As a result, a total of about 4.6 million molecules were retained and then docked with Glide (Standard Precision mode with default parameters). Molecules with docking scores below -5.0 cutoff were further docked with the FRED program, OEDocking (up to 500 conformers were generated for each molecule and were docked using FRED with default parameters), and the corresponding RMSD values were calculated for the top poses. All the molecules with an RMSD \( \leq 3\text{A} \) were retained and were docked again using ICM. For the poses predicted by ICM (default parameters), RMSD values against Glide were calculated and only the molecules with RMSD \( \leq 3\text{A} \) were retained.

Within this set, predicted docking pKi was calculated for each molecule using a custom MOE SVL script. Other properties like ADMET (absorption, distribution, metabolism, excretion, toxicity) and pharmacokinetics predictions were also calculated by using computational programs such as ADMET Predictor, FAF-Drugs and Quantitative Estimate of Drug-likeness (QED). In the next step, a consensus scoring method was used and Molecules with total consensus scores greater than 5 were retained and then clustered together to remove the similar compounds (70% similarity). Finally, a total of 52 compounds were chosen for experimental validation.

### 3.2.3 In vitro screening

All compounds were dissolved in DMSO as a 50mM stock solution and diluted into treating concentration with growth medium (RPMI 1640 (Hyclone, GE) containing 10% FBS (Thermofisher Scientific) and 1X Anti-anti (Antibiotic-Antimycotic, Thermofisher Scientific)). Suspension cell lines Hut78, Jurkat, K562, and U937 were purchased from ATCC. SZ4 and Mac2A cell lines were generous gifts from Dr. Ivan Litvinov\(^{128}\). The cell was cultured in the growth medium and collected at the logarithmic growth phase (about 5-10x10^5/ml). Cells were
seeded into 96 well culture plates (Nunc, Thermofisher Scientific) with 10^4/well. Cells were cultured with various concentrations of the testing compounds in 0.2% DMSO or DMSO control only (as 0 µM) in growth medium for 64-68 hours in incubator containing 5% CO2 at 37 degrees. Viability assay was performed using CellTiter-Blue® Cell Viability Assay (Promega), and a fluorescent signal (579Ex/584Em) was recorded after 2 hours and 4 hours of incubation time using Glomax Multi Detection System (Promega).

All treatments were with triplicates, and the final value was calculated as the mean of the 3 datasets after subtracting medium only background. Net fluorescent signals at various concentrations were then compared to the DMSO only control and calculated as the percentage of the surviving population. IC50 was then estimated using an online IC50 calculator (AAT Bioquest).

RNA expression of TOX was measured by dye intercalated Realtime PCR method described previously^97. RNA was extracted from cells using RNAeasy purification kit (Qiagen), and cDNA template was produced using SuperScript™ VILO™ cDNA Synthesis Kit (Invitrogen, Thermofisher Scientific). Gene expression levels were expressed as mRNA copies per 1000 glyceraldehyde-3-phosphate dehydrogenase (GAPDH) copies by standardizing to internal housekeeping gene GAPDH. The primers used for real-time measurement are as following: GAPDH forward AAGATCATCAGCAATGCCTCC, GAPDH reverse TGGACTGTGGTCATGAGTCCTT, TOX forward GTGCAGAAATCCTCCCCAC, TOX reverse TTTGTCCCTCTGCATGCCC.
3.3 Results

3.3.1 Druggability assessment of the TOX HMG-box domain

For in silico screening of small molecules we used the available NMR structure of TOX (Figure 5) from the PDB database (ID: 2CO9). We first performed a pilot study by screening about 200,000 small molecules (having drug-like properties from ZINC database) using the “blind docking” setup of docking software GLIDE. This was done to both investigate the druggability of the DNA-binding domain of TOX as well as to identify the binding hotspots of small molecules on the protein surface. Some of the hotspots are near the DNA interface on the HMG-box domain which can be seen from Figure 6.

![Figure 5. Protein structural templates for the HMG-box domain of TOX.](image)

(a) Mouse TOX HMG Box domain (PDB: 2CO9)
(b) Human TFAM HMG Box domain with DNA (PDB: 3TMM)
(c) Superposition of 2CO9 and 3TMM determines the protein-DNA interface on TOX

Figure 5. Protein structural templates for the HMG-box domain of TOX. (a) An NMR structure of mouse TOX protein (PDB ID: 2CO9) has been identified as the best structural template, with 100% sequence similarity across the 87 amino acids of the HMG-box domain, compared to the human TOX protein. (b, c) By superimposing the 2CO9 structure (orange ribbons) onto the HMG-box protein TFAM (pink ribbons, PDB ID: 3TMM, 46% sequence similarity to human TOX, 3.8A RMSD) in complex with DNA (green ribbons), the TOX-DNA interface was determined.
3.3.2 Large-scale in silico screening

We then performed a large scale virtual screening of about 7.6 million molecules with drug-like properties taken from the ZINC database, details of which are mentioned in the method.
section. Finally, 118 compounds were selected based on the top consensus scores for in vitro testing, 66 compounds with a molecular weight greater or equal to 350 Dalton, and 52 compounds with molecular weight less than 350 Dalton.

### 3.3.3 In vitro experimental validation

All the compounds including 22 compounds from initial screening and 118 compounds from large-scale screening were experimentally tested by using TOX dependent CTCL cells (Hut78 cells) at 10 µM and 100 µM concentrations. Out of all 140 compounds, 18 compounds showed concentration-dependent inhibition of cell viability in Hut78 cells, their IC50 values were determined in 3 TOX-high/dependent CTCL cell lines (Hut78, SZ4, Jurkat), and 3 TOX-low/independent cell lines (K562, U937, Mac2A) (Table 2). The IC50 values for these 18 compounds were lower in TOX-high cell line compared to TOX-low cell lines (i.e. the molecules are selective towards TOX-high cell line, Table 2). Several of these small molecule inhibitors (SMIs), such as 190444 and 190414, have IC50 values in the range of 10~20 µM, more active than the hit compound 190010 that were identified from the initial in silico screen. As further illustrated in Figure 7, compounds 190444, 190414, 190447 and 190441 inhibited cell viability of the TOX-high cells (Hut78, Jurkat) selectively, compared to the TOX-low cells (K562). In addition, Figure 8 shows that compounds 190444, 190414 and 190441 increased the expression of SMAD3, which is normally suppressed by TOX$^{97}$.

Table 2. Top candidates for TOX small molecule inhibitors (SMIs).

<table>
<thead>
<tr>
<th>Compound VPC-ID</th>
<th>Chemical Structure</th>
<th>Average IC50 (µM) (TOX-High cells)</th>
<th>Average IC50 (µM) (TOX-Low cells)</th>
<th>TOX-selectivity index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>Molecular Structure</td>
<td>33.26</td>
<td>56.29</td>
<td>1.69</td>
</tr>
<tr>
<td>-------</td>
<td>--------------------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>190343</td>
<td><img src="image1.png" alt="Molecule 190343" /></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>190341</td>
<td><img src="image2.png" alt="Molecule 190341" /></td>
<td>52.44</td>
<td>85.82</td>
<td>1.64</td>
</tr>
<tr>
<td>190325</td>
<td><img src="image3.png" alt="Molecule 190325" /></td>
<td>39.77</td>
<td>64.63</td>
<td>1.63</td>
</tr>
<tr>
<td>190323</td>
<td><img src="image4.png" alt="Molecule 190323" /></td>
<td>43.79</td>
<td>68.68</td>
<td>1.57</td>
</tr>
<tr>
<td>190339</td>
<td><img src="image5.png" alt="Molecule 190339" /></td>
<td>34.02</td>
<td>51.83</td>
<td>1.52</td>
</tr>
<tr>
<td>190322</td>
<td><img src="image6.png" alt="Molecule 190322" /></td>
<td>24.64</td>
<td>36.19</td>
<td>1.47</td>
</tr>
<tr>
<td>190349</td>
<td><img src="image7.png" alt="Molecule 190349" /></td>
<td>52.88</td>
<td>70.57</td>
<td>1.33</td>
</tr>
<tr>
<td>190301</td>
<td><img src="image8.png" alt="Molecule 190301" /></td>
<td>38.75</td>
<td>47.29</td>
<td>1.22</td>
</tr>
<tr>
<td>190354</td>
<td><img src="image9.png" alt="Molecule 190354" /></td>
<td>58.59</td>
<td>66.12</td>
<td>1.13</td>
</tr>
</tbody>
</table>
Average IC50 values of cell viability from 3 TOX-high/dependent CTCL cell lines (Hut78, SZ4, Jurkat).

Average IC50 values of cell viability from 3 TOX-low/independent lymphoid cell lines (K562, U937, Mac2A).

TOX-selectivity index = Average IC50 (TOX-Low cells) / Average IC50 (TOX-High cells).

Figure 7. Viability curves of TOX-high and TOX-low expressing cells. Cells were treated with various concentrations of compounds, (a) 190444, (b) 190414, (c) 190447 and (d) 190441, for 72 hours in 37°C incubator with 5% CO2. Viability was measured by CellTiter-Blue® assay and compared to the DMSO control as described in Methods and Materials. Jurkat and Hut78 cells (solid line) are the TOX-high expressing cell lines, while K562 (dotted line) is the TOX-low expressing cell line. (Produced in collaboration with Dr. Zhou’s lab)
Figure 8. Viability curves of TOX-high and TOX-low expressing cells. Cells were treated with various concentrations of compounds, (a) 190444, (b) 190414, (c) 190447 and (d) 190441, for 72 hours in 37°C incubator with 5% CO2. Viability was measured by CellTiter-Blue® assay and compared to the DMSO control as described in Methods and Materials. Jurkat and Hut78 cells (solid line) are the TOX-high expressing cell lines, while K562 (dotted line) is the TOX-low expressing cell line. (Produced in collaboration with Dr. Zhou’s lab)

3.4 Discussion and future direction

TOX has been identified as a promising drug target for CTCL therapies based on the previous study\(^99\) and since there are no SMI’s for TOX, we used this unmet opportunity to discover new TOX inhibitors. Using our established in house pipeline (chapter 1) and in vitro experiments we discovered 18 SMIs for TOX, which can inhibit the viability of TOX-high/dependent cells with micro-molar IC50 and up to 4 folds’ selectivity (Table 2). As illustrated in Figure 9, compounds
190444, 190414, 190447 and 190441 can bind at the hot spots located in close proximity to the protein-DNA interface on the HMG-box domain of TOX. These SMIs interact with TOX protein residues including Gln262, Pro264, Arg273, Lys313, Glu320, Gln324 through hydrogen-bond and hydrophobic interactions, corresponding well to the hot spots as identified from the druggability assessment (Figures 6 and 9).

Given the proximity of binding, it is likely that the small molecules can interfere with TOX-DNA interactions and inhibiting the activity of TOX. This hypothesis is partially supported by the experimental results, compounds 190444, 190414 and 190441 increased the expression of SMAD3, which is normally suppressed by TOX (Figure 8). To establish the hypothesis further, additional experiments, including direct binding, DNA-competition, and luciferase reporter assays, are required. Future development of drug candidates inhibiting TOX-DNA interactions can follow previous studies where SMIs have been successfully developed via CADD to target the DNA-binding domains of other cancer drug targets such as AR, ERG, and MYC.

It takes more than 3 billion dollars and 10 years for a drug to reach patients from a laboratory. We demonstrated how virtual screening can significantly speed up time as well as save a lot of money during the drug discovery phase. 18 compounds were found active out of 140 in vitro tested compounds, achieving a hit rate of 13% (18/140) which is a lot higher than <1% from conventional, experimental high-throughput screening without any computational guidance\textsuperscript{29, 30}. These 18 hit compounds provide the foundation where more potent TOX-SMIs can be developed through 2D/3D similarity searches (ligand-based screening) of chemical analogs against the entire ZINC database, which has grown exponentially from 700,000 compounds in 2005\textsuperscript{126} to over 1 billion molecules in 2019\textsuperscript{32}. Until now we were just using <10 million molecules (7.6 million in case of TOX) i.e. <1% of entire ZINC database and throwing away remaining >99% due to time
It has been shown before\textsuperscript{42} that larger library contains molecules that are better suited for a given receptor structure than can be found in smaller libraries. This ZINC database of chemicals provides a tremendous opportunity for TOX drug discovery where virtual screening by molecular docking can be expanded from the initial 7.6 million to all of the 1 billion molecules. In the next section, we are going to talk about “Progressive Docking 2.0” that trains a machine learning model to efficiently predict binding scores based on chemical structures; thus compute-intensive docking only needs to be performed on a subset of molecules pre-calculated as good target-binding candidates\textsuperscript{67}. We anticipate applying such a PD2.0 algorithm (with >50x speed-up) to virtually screen 1 billion molecules against TOX.

Both of these approaches, similarity search, and PD2.0, can greatly expand our current collection of TOX hit compounds and enable us to build Quantitative Structure-Activity Relationship (QSAR) models\textsuperscript{129} that can guide future development for the next generation of potent and selective TOX drug candidates.
Figure 9. TOX small molecule inhibitors (SMIs) bind at the hot spots on the protein-DNA interface. Docking poses are shown for compounds (a) 190444, (b) 190414, (c) 190447 and (d) 190441 on the HMG-box domain of TOX. For each panel: left) molecular surface representation, SMI in cyan, DNA in green ribbons, hot spots as red surface patches; right) in-depth molecular interactions between SMI (cyan) and protein residues of TOX (orange), hydrogen-bonds in red dotted line, hydrophobic interactions in green dotted line.
Chapter 4: Accelerating docking using progressive docking 2.0

4.1 Introduction

VS of ultra-large chemical libraries is an emerging approach in CADD which recently was proved successful in identifying novel chemotypes with high potency. Despite the great potential for drug discovery, current methods are limited to screen ~100M molecules and lag behind the unceasing growth of databases such as ZINC15 (1.3 billion of structures) or Enamine REAL (700 million structures). Thus, there is a pressing need to accelerate the standard docking process. We came across the very same need when we targeted TOX protein using standard docking protocols, which limited our VS campaign to less than 1% of the ZINC database (7.6M compounds) due to the computational cost associated with larger-scale screening. In this chapter, we report the development and validation of a new deep learning-based method called “Progressive Docking 2.0” (PD2.0) which allows up to 65 folds docking speed increase on a variety of drug targets.

Components of the developed PD2.0 protocol are shown in Figure 10. The first step of the protocol is to obtain the SMILES and calculate molecular features (fingerprints) for all the small molecules of the library to be screened. In the first PD2.0 iteration, a fixed number of molecules (e.g. 3 million structures) are randomly sampled from the database and docked using GLIDE or FRED. Docking scores obtained are then converted to binary values (0=low binding affinity, 1=high binding affinity) based on a docking score threshold (details of which are explained later). In this way, we convert a regression problem into a classification problem. The data is then divided into training, testing and validation sets. QSAR models are built and evaluated, and the best model is used to predict the class (0/1) for each of the molecules in the entire dataset (e.g. 1B molecules). For the successive iterations, the same number of molecules are randomly sampled from the set of molecules predicted as positives during the previous iteration, and docked. These molecules are
then added to the training dataset from the previous iteration. The validation and testing set remains the same as the first iteration for all the iterations. Thus, the distribution of scores in validation and testing sets are conserved while the training set distribution might change. QSAR models are again generated and evaluated, and the best model is used to predict the classes for all the compounds of the original database. This process is repeated until the number of positive predictions returned by PD2.0 (e.g. 5-10 million good binding candidates) can be docked in a reasonable amount of time, given the computational resources available.

**Figure 10.** Different steps involved in PD2.0. Features are generated from the SMILES of the small molecule in the library and stored in a feature database. For the first iteration, a small subset of molecules (e.g. 3 million) is randomly sampled and divided into training, validation and testing set, and docked against the protein target. A score cut-off value is chosen to convert the problem from regression to classification. A QSAR model is built and used to predict the classes for all the molecules of the database (1=good and 0=bad). From the next iteration
onwards, the same number of molecules (3 million) are sampled only from the positive molecules (predicted to be good binders) of the previous iteration, while ‘bad’ molecules are discarded. The validation and testing set from the first iteration remain the same for all the iterations, and the existing training set is enriched with the newly sampled molecules. All the other steps remain the same as the first iteration. This process is repeated multiple times until the total number of positively predicted molecules is a reasonable number that can be docked.

4.2 Methods

4.2.1 Molecule fingerprints

Chemical fingerprints\textsuperscript{130} is an easier molecular format to be interpreted by a computer, compared to SMILES\textsuperscript{131}. A fingerprint is a series of binary bits that can vary from 100-4k in size. Typically, a kernel is applied to generate a bit vector or count vectoring order to get molecule features, hash them and determine bits based on the hashing.

MACCS\textsuperscript{132} fingerprints rely on a binary representation of the presence and counting of specific atoms and rings in a molecule. For instance, one of the MACCS bits assesses if the molecule has more than 2 nitrogens by having a value of 1 (true) or 0 (false). Another bit will assume a value of 1 or 0 depending on the presence of more than 2 aromatic rings, and so on. MACCS fingerprint generated with RDKIT\textsuperscript{133} provides a 166-bit vector.

Another type of fingerprints is Morgan\textsuperscript{130} fingerprints, which are 2D representations of molecules derived from the Extended Connectivity FingerPrint\textsuperscript{134} (ECFP). These fingerprints account for internal graph connectivity as they evaluate the neighborhood for each atom of the structure. All possible paths within a given radius through an atom are obtained and each path is hashed into a number based on the number of specified bits. Higher radius will result in bigger fragments being encoded, thus more bits will allow Morgan fingerprints to be more selective among the molecules.
Pharmacophore fingerprints\textsuperscript{135} consider any pair of pharmacophore features, such as hydrophobic atoms or groups, hydrogen bond acceptors, and hydrogen bond donors. Each pair is assigned a number of bits corresponding to non-overlapping ranges of topological distances. When evaluating the pharmacophore fingerprint of a given molecule, bits corresponding to feature pairs present in the structure at specific distance ranges are set to \textsuperscript{136}.

4.2.2 Molecular docking

All the complexes from the PDB repository were prepared and optimized using Protein Preparation Wizard\textsuperscript{137} of Schrödinger suite. Receptors were prepared for docking using the Make Receptor utility from OpenEye and Schrödinger’s Grid Preparation Wizard. Docking was performed using OpenEye’s FRED\textsuperscript{87} docking program and Schrödinger’s GLIDE\textsuperscript{86} program.

4.2.3 Clustering

In clustering, objects that are similar to each other are grouped together while dissimilar objects are divided into different groups. Extracting representative objects from each cluster instead of random sampling ensures to cover the diverse object space more accurately.

4.2.3.1 K-means clustering

In a K-means\textsuperscript{9} clustering algorithm, a data point is associated with the cluster (K groups) with the most similar centroids (means) to it. Thus, for each data point $x$, the corresponding cluster will be the one which satisfies:
\[
\arg\min_{c_i \in C} \text{similarity}(c_i, x) \tag{1}
\]

Here \(c_i\) is the collection of centroids in set \(C\), and the similarity measure is usually the squared \(L_2\) (Euclidean) distance. After each iteration where all data points assigned to respective clusters, the centroids are recalculated:

\[
c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i \tag{2}
\]

where \(S_i\) is the set of data points of the \(i^{th}\) cluster. The process is repeated until convergence, i.e. all the samples are assigned to the same cluster in consecutive iterations. As K-means algorithms converge to local optima, it is important to run multiple initializations and choose the best one based on the residual sum of squares (RSS) objective function:

\[
RSS_k = \sum_{x \in \omega_k} |x - \mu(\omega_k)|^2 \tag{3}
\]

The time complexity of K-means clustering using the squared \(L_2\) distance as similarity measure is \(O(nkId)\), where \(n\) stands for the number of data points, \(k\) stands for the number of centroids, \(I\) stand for the number of iterations and \(d\) stand for the dimensionality of data. In our PD2.0 pipeline, we deal with more than 1 billion molecules, each one having a Morgan fingerprint of 1024 bits. Thus, running standard K-means clustering for any realistic number of clusters was computationally infeasible. Similarly, searching for an optimal number of cluster was infeasible since it would have taken an extremely large amount of time just for one run.

**4.2.3.2 Mini-batch K-Means clustering**

Mini-batch K-means\textsuperscript{138} is an approximation of K-means clustering, which uses a mini-batch of fixed size for updating the cluster centers at each iteration. The method is fast and suitable
for large data sets, and it also requires less memory space. A detailed trade between accuracy and speed is described in detail in the paper by Feizollah et al\textsuperscript{139}. The major takeaway is that as the number of clusters increases the speed gets an exponential boost while at the same time clusters start to diverge from K-means clusters. This issue becomes more prominent as the data set size increases.

4.2.4 Logistic regression

Binary logistic regression\textsuperscript{4} is logistic regression where the algorithm is used to assign observations to two discrete classes. Logistic regression is based on probability; it uses a sigmoid function/logistic function (Equation 4) to map the probability of observation between 0 and 1. The decision boundary generated by logistic regression is linear in nature and is decided by a cut-off value chosen based on some accuracy metrics.

$$Sigmoid \; \sigma(z) = \frac{1}{1 + e^{-z}} \quad (4)$$

The objective function/cost function used for logistic regression is a convex function called logistic loss/cross-entropy loss\textsuperscript{140} (Equation 5). This loss is minimized to obtain the global optimum solution.

$$J(\theta) = -\frac{1}{m} \sum \left[ y^i \log \left( h_\theta(x(i)) \right) + (1 - y^i) \log \left( 1 - h_\theta(x(i)) \right) \right] \quad (5)$$

4.2.5 Support Vector Machine

The goal of Support Vector Machine (SVM) is to find the hyperplane which separates the N-dimensional space into distinct classes by maximizing the distances (margins) between the data points belonging to different classes\textsuperscript{6}. SVM uses support vectors i.e. the data points of each class
which is closest to the hyperplane. The presence of margin is a form of reinforcement towards classifying future data with higher confidence. The loss function representing the margins is given by hinge loss with an L2 regularizer\textsuperscript{140} (Equation 6), which is minimized. The defined loss is a convex function thus the solution obtained is a global optimum.

\[
J(\theta) = \lambda ||w||^2 + \sum_{i=1}^{n} \left(1 - y_i < x_i, w > \right)_+
\]

(Equation 6)

The decision boundary obtained is again linear in nature but non-linear decision boundaries can be obtained via the kernel trick\textsuperscript{141}. Here a kernel function is used to map the non-linear separable input data to higher dimensional separable features. The solution is then mapped back to the non-linear separable space. The type of kernel which is most used includes Radial Basis Function\textsuperscript{142} (RBF) kernel (Equation 7) and polynomial kernel\textsuperscript{141} (Equation 8).

\[
K(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right)
\]

(Equation 7)

\[
K(x, x') = (x^t y + c)^d
\]

(Equation 8)

4.2.6 Random forest

Random forest\textsuperscript{5} (RF) consists of an ensemble of decision trees each built using bootstrap sample and a subset of features. Since decision trees usually have high variance since they overfit the training data, this overfitting can be reduced by using multiple decision trees generating independent mistakes, and making decisions based on the majority. Trees making independent mistakes can be obtained by using bootstrap samples i.e. multiple training sets generated with replacement and only using a subset of features for each tree. This strategy not only ensures
diversity among trees different but also reduces the training time since not all the features are used for training. All the decision trees can be independently trained in parallel. During the final prediction stage, results from each tree are obtained and a decision is made based on the majority.

4.2.7 Feedforward neural network

The neural network approach was initially inspired by how human neurons work. A typical NN consists of multiple neurons (nodes) arranged in multiple layers, with all nodes in adjacent layers having connections to each other. Layers are divided into an input layer (data features), output layer (predictions) and hidden layers (intermediate layers connecting an input to output). Each node in the hidden layer transforms the input features to generate more refined features (activation) by taking a linear combination of all the input features and applying non-linear transformations like the rectified linear unit to it:

\[
\hat{y}_i = w^t \left( \frac{1}{m} \sum_{l=1}^{m} h(w^l x_i) \right)
\]

Finally, these refined features are sent to an output layer to either perform a regression task or classification task. The loss function is different for regression (mean squared error, Equation 10) and classification tasks (cross-entropy error, Equation 5), it is non-convex in nature, due to the introduction of non-linearity after each layer, and it gives a local optimum solution.

\[
J(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2
\]
4.2.8 Evaluation metrics

Evaluation metrics are used to test the performance of different models and select the best one. There is a variety of evaluation metrics available for regression and classification tasks. For our classification models, we have used Receiver Operating Characteristic (ROC) curve, precision-recall, and enrichment, which are discussed below.

ROC curve is a plot of true positive rate (TPR or sensitivity or recall, Equation 11) against false positive rate (FPR or fall-out, Equation 12) at multiple threshold settings.

\[
TPR = \frac{TP}{TP + FN} \quad (11)
\]

\[
FPR = \frac{FP}{FP + TN} \quad (12)
\]

TPR and FPR both increase simultaneously but ideally TPR should increase a lot faster than FPR at the beginning and plateau at 1, while FPR should slowly increase to 1. The area under the ROC curve (AUC) is the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Thus, a model with higher AUC works better than a model with a lower value.

In a classification problem, task precision is defined as the total number of true positives (TP) divided by the total number of elements model predicted to be positive (TP + false negatives, FN) (Equation 13). Precision represents what fraction of predictions is actually correct for a specific class.

\[
Precision = \frac{TP}{TP + FN} \quad (13)
\]
On the other hand, recall is the total number of TP divided by total elements that actually belong to the positive class (TP + false positives, FP) (Equation 14), and it represents the fraction of elements of a class correctly classified by the model.

\[
\text{Recall} = \frac{TP}{TP + FP} \tag{14}
\]

Thus, both high recall, as well as high precision score, are desirable in a model.

We defined two types of enrichment: 1) full predicted database enrichment (FPDE) and 2) top n enrichment. FPDE (Equation 16) is defined as the ratio of the fraction of positive class which is predicted positive (precision, Equation 13) and the fraction of positive class in the dataset (random precision, Equation 15). FPDE indicates the TP enrichment of the dataset after a model iteration compared to the initial dataset. Top n enrichment (Equation 17) is defined as a ratio of the total number of positive class elements in top n data points predicted by the model (where the rank is established by the DNN probabilities) and n data points randomly selected from the same dataset.

\[
\text{Random Precision} = \frac{\text{\#positive in dataset}}{\text{\#molecules in dataset}} \tag{15}
\]

\[
\text{FPDE} = \frac{\text{Model Precision}}{\text{Random Precision}} \tag{16}
\]

\[
\text{Top n enrichment} = \frac{\text{\#positive in models top n ranking}}{\text{\#positive in random n}} \tag{17}
\]

4.2.9 Conformal prediction

Conformal prediction is a method of prediction that can be added to any classification or regression algorithms like logistic regression, SVM or Neural network. It was invented by Vovk et al\textsuperscript{143}. The main advantage of using it over the threshold-based approaches is that users can define
a confidence interval i.e. given a confidence threshold of $\epsilon$ it is guaranteed that the actual label has a recall value of 90% (we only miss 10% of the actual compounds). For a binary classification model, four classes are predicted: class A, class B, both class A, B and the empty set. For our application we set all the predictions to be the negative class/class B if not predicted to be just class A. We have used Inductive Conformal Prediction (ICP) over transductive approach since the transductive method is very computationally expensive.

For transductive conformal prediction, a model is built on a training set and then a calibration set (validation/test set) is used to calculate the nonconformity score. This score captures the information on how different the calibration/validation/testing set is from the training set. Once the calibration is done, for predicting on a new dataset, the p-value for all the classes (in our case binary) is calculated and the classes for which the p-value is greater than a predefined threshold $\epsilon$ are assigned to the data point. For example, for a binary classification problem let the p-value for class 0 and 1 be 0.05 and 0.3 respectively. Then given $\epsilon$ to be 0.1 (90% confident or 10% error rate) the data point will be assigned to class 1. With this method, one can make sure that the right label is present in the prediction set with a probability of 1- $\epsilon$.

4.3 Results and discussion

4.3.1 Molecular featurization

To evaluate the effect of different molecular fingerprints to model performances, a random subset of 1 million molecules from ZINC 3D database$^{32}$ (570 million molecules) was divided into training, testing and validation sets, and docked with FRED$^{87}$. MACCS, Morgan, and pharmacophore fingerprints were calculated for all the 1 million molecules. Models were built for each fingerprint using the same hyperparameters, and different top n enrichment values such as
top 10, top 100, top 1000, top 10000 and FPDE at a fixed value of 90% for recall were calculated. The fingerprint with the highest values in the majority of enrichment scores was selected for PD2.0. Enrichment values are shown in Table 3. The best scoring fingerprint was Morgan with a dimension of 1024 and radius 2.

Table 3. Different enrichment values for MACCS, Morgan, and Pharmacophore fingerprints.

<table>
<thead>
<tr>
<th>Fingerprint</th>
<th>Top 10 enrichment</th>
<th>Top 100 enrichment</th>
<th>Top 1000 enrichment</th>
<th>Top 10000 enrichment</th>
<th>FPDE (90% recall)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MACCS</td>
<td>71</td>
<td>50</td>
<td>40</td>
<td>23</td>
<td>3.7</td>
</tr>
<tr>
<td>Morgan_512_2</td>
<td>71</td>
<td>92</td>
<td>59</td>
<td>30</td>
<td>4.1</td>
</tr>
<tr>
<td>Morgan_512_3</td>
<td>141</td>
<td>89</td>
<td>58</td>
<td>27</td>
<td>3.5</td>
</tr>
<tr>
<td>Morgan_1024_2</td>
<td>164</td>
<td>103</td>
<td>71</td>
<td>31</td>
<td>4.2</td>
</tr>
<tr>
<td>Morgan_1024_3</td>
<td>70</td>
<td>92</td>
<td>64</td>
<td>29</td>
<td>3.8</td>
</tr>
<tr>
<td>Morgan_2048_2</td>
<td>141</td>
<td>82</td>
<td>69</td>
<td>31</td>
<td>4.3</td>
</tr>
<tr>
<td>Morgan_2048_3</td>
<td>94</td>
<td>85</td>
<td>65</td>
<td>29</td>
<td>3.7</td>
</tr>
<tr>
<td>Pharmacophore</td>
<td>0</td>
<td>14</td>
<td>22</td>
<td>14</td>
<td>1.7</td>
</tr>
</tbody>
</table>

4.3.2 Database sampling

To evaluate different sampling techniques, 3 million molecules (test dataset) were randomly sampled from the ZINC database (1 billion molecules) and docked against ER-AF2 using GLIDE. The molecules were then clustered using mini-batch K-means clustering into 10, 100, 500, 1000, 2000, 3000, 4000 and 30000 clusters. We then performed a single iteration of PD2.0 by sampling from clusters (where the number of molecules sampled from each cluster was proportional to the size of cluster) as well as by performing random sampling on the whole database. For each cluster size, we sampled training, validation, and testing set, each consisting of...
the same number of molecules. We used the same hyperparameters for building the models. We tested different sample sizes and compared the stability, generalizability, and accuracy of the two methods. The criteria for determining the stability of a model was to retrain the model multiple times (5 times) by resampling each time and analyzing how stable the FPDE values were for the test dataset. Generalizability was determined based on how the validation precision (criterion 2) and recall (criterion 1) transferred to the test dataset and accuracy was the measure of FPDE (at 90% recall) for each of the models.

We observed that model stability, as well as accuracy, was similar irrespective of sampling from the cluster or from the random at all the sample sizes (Figure 12). The model generalizability was also very similar. The plots for the ratio of test recall to validation recall and test precision to validation precision are shown in Figures 13 and 14.

To assess how cluster compactness related to the docking scores, we randomly sampled and docked 10000 molecules. We then performed clustering based on multiple Tanimoto similarities\textsuperscript{144}. We observed that at higher cutoff (0.8-0.9) molecules within each cluster were similar but too many clusters were generated, including many singletons. When cutoff was set to lower values of 0.5 or less, we observed fewer clusters but the docking scores distribution within each cluster was the same as the distribution of the whole dataset (Figure 11).

After all these different experiments we could safely conclude that clustering does not provide any significant advantage over random sampling given the heavy computational cost associated with it.
Figure 11. Docking score distribution of the entire dataset versus the docking score distribution of the largest cluster when a Tanimoto cutoff of 0.5 was used. The distributions are overlapping hence all the molecules within the cluster have the same distribution of scores as that of random sampling.
Figure 12. Accuracy and stability of the QSAR model obtained from samples of clustered and unclustered data. Each figure is a plot of FPDE (at 90% recall) against the number of clusters for different sample sizes ranging from 3000 to 480000. The mean, as well as the standard deviation, does not show any trend for any sampling technique.
Figure 13. Generalizability (criterion 1) of the QSAR model from samples of clustered and unclustered data. Each figure is a plot of the ratio between test and validation recall values against the number of clusters for different sample sizes ranging from 3000 to 480000. The mean, as well as the standard deviation, does not show any trend for any sampling technique.
Figure 14. Generalizability (criterion 2) of the QSAR model from samples of clustered and unclustered data. Each figure is a plot of the ratio between test and validation precision values against the number of clusters for different sample sizes ranging from 3000 to 480000. The mean, as well as the standard deviation, does not show any trend for any sampling technique.
4.3.3 QSAR model for docking scores

We tested four different machine learning models: RF, Logistic Regression, SVM and DNN to select the best-suited one for PD training. 3 million molecules were sampled and docked against the ER-AF2 site to obtain an evaluation dataset for model testing. Furthermore, a basic grid search was performed to select the best hyperparameters. Enrichment values for all the models are reported in Table 4. SVM took an extraordinary amount of time for training compared to other approaches (>10 hours), therefore the method was discarded.

DNN generated the highest enrichment values, as we expected given the large size of the dataset. It should also be noted that docking is the rate-determining step in the PD2.0 pipeline. Thus, the small enrichment improvement of DNN compared to RF would have resulted in a significantly lower number of molecules left to be docked after the final iteration, resulting in a notable speed boost. From this analysis, we concluded that DNN was the best model for our application and used it for all the further modeling.

The general architecture for the DNN used for PD2.0 consists of 1-4 hidden layers, each layer consisting of the same number of neurons. A dropout layer was present after each fully connected layer. The number of hidden layers, number of neurons and dropout frequency were the hyperparameters of the model. The scheme is shown in Figure 15.
Table 4. Comparison of different QSAR models: Feedforward Neural Network (DNN), Random forest (RF), and logistic regression. DNN provides the best scores for all 5 enrichment values (top 10, top 100, top 1000, top 10000, FPDE at 90% recall).

<table>
<thead>
<tr>
<th>Method</th>
<th>Top 10 enrichment</th>
<th>Top 100 enrichment</th>
<th>Top 1000 enrichment</th>
<th>Top 10000 enrichment</th>
<th>FPDE (90% recall)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN</td>
<td>164</td>
<td>103</td>
<td>71</td>
<td>31</td>
<td>4.2</td>
</tr>
<tr>
<td>RF</td>
<td>117</td>
<td>75</td>
<td>59</td>
<td>28</td>
<td>4.1</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>47</td>
<td>49</td>
<td>46</td>
<td>24</td>
<td>3.7</td>
</tr>
</tbody>
</table>

Figure 15. The architecture of the DNN used for the QSAR model. One block is the combination of one fully connected layer and one dropout layer. The number of blocks, number of neurons and dropout frequency are the hyperparameters of the model. The input layer consists of 1024 neurons taking values from Morgan radius 2 and 1024 bits fingerprints. The output layer consists of two neurons as the problem is of binary classification.
4.3.4 Progressive docking 2.0

We validated PD2.0 on a wide range of protein targets. We screened 1.3 billion small molecules from ZINC15 against ER-AF2 using GLIDE. To our knowledge, this is the first successful virtual screening performed on more than 1 billion molecules. We further evaluated PD2.0 robustness by docking 570M molecules on 12 drug targets belonging to different protein families (Table 1), using FRED. Thus, we conclude that PD2.0 is a suitable technique for large-scale virtual screening which can be used for a variety of protein targets and docking programs.

4.3.4.1 Training data for PD2.0

To determine the optimal number of molecules to sample and dock at each PD iteration, we plotted the training sample size versus FPDE with recall fixed at 90% (Figure 16): larger training sample sizes lead to higher FPDE values.

Also, we observed a decrease in the standard deviation with the increase of the sample sizes, indicating an increase of the model stability (Figure 17). Consequently, we decided to sample at least 1 million molecules for each training iteration. In future applications of PD2.0, the training sample size can be user-determined based on the available computational resources. In order to convert the continuous docking score values to a classification problem, we chose a cut-off value based on a required number of positive molecules (molecules with docking score lower than a cutoff value) to be observed in the validation set after the last iteration. At each i\text{th} iteration, the cut-off was chosen as the value that returns a rescaled number of positive molecules observed in the validation set equal to:
\[ N_{\text{pos},i} = N_{\text{pos},\text{last}} \cdot \frac{\text{mol}_{i-1}}{\text{mol}_{\text{last}}} \]  

where \( N_{\text{pos},\text{last}} \) is the minimum number of positive molecules in the validation set after the last iteration (200 in our case), \( \text{mol}_{i-1} \) is the number of molecules left after the \( i^{\text{th}} \)-1 iteration and \( \text{mol}_{\text{last}} \) is the minimum number of molecules to be left after the last iteration (15 million in our case).

The best hyperparameter combination was selected based on the model with a recall of 90% and the highest precision or FPDE. A decrease of the docking score cut-off after each iteration was expected and observed, indicating that the definition of positive molecules (i.e. good binders) improves after each iteration. In order to deal with a highly imbalanced classification dataset (ranging from 0.1%-2% positives), we have used both oversampling of the minority class as well as class weights, where the class weight and the oversampling ratio were the hyperparameters of the model.
Figure 16. The plot of FPDE against training dataset size for random sampling. FPDE increases with the increasing of the dataset size.

![Figure 16](image)

Figure 17. Generalizability (criterion 1)/recall ration for random sampling. (a) The recall ratio i.e. ratio of validation recall and test recall becomes more stable (smaller spread) with bigger sample size. (b) The standard deviation of the recall ratio goes down with a larger sample size.

4.3.4.2 PD2.0 on Estrogen Receptor Activation Function 2 (ER-AF2)

The ER-AF2 site was obtained from the Protein Data Bank (PDB 3UUD) and prepared using SCHRODINGER 2016 package as described in Chapter 1. All the 1.3 billion SMILES of the ZINC15 2D database were downloaded and Morgan fingerprints with radius 2 and 1024 bits were calculated. We then applied our PD2.0 pipeline to accelerate Glide docking on the AF2 site.
4.3.4.2.1 Threshold-based prediction

All the molecules with a model-predicted probability above a threshold were allotted to the positive class. We determined this threshold as the value at which the model recall was 90%, as discussed previously.

A new DNN model was trained at each PD iteration. A basic hyperparameter grid search was performed and the best model was chosen for prediction on the entire dataset (1.3 billion molecules). Table 5 shows the resulting model statistics of the best model at each iteration.

Table 5. Different model statistics for threshold-based prediction after each iteration for ER-AF2. ROC-AUC, different enrichment values and recall values for the best model are reported. We observed an increase in all enrichment values after each iteration. The model cutoff/definition of good molecules improved after each iteration as well.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Model cutoff</th>
<th>Model ROC AUC</th>
<th>Top 10</th>
<th>Top 100</th>
<th>Top 1000</th>
<th>FPDE</th>
<th>Model Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-5.24</td>
<td>0.923</td>
<td>39</td>
<td>28</td>
<td>26</td>
<td>4.14</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>-5.84</td>
<td>0.97</td>
<td>218</td>
<td>144</td>
<td>104</td>
<td>11.11</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>-6.95</td>
<td>0.994</td>
<td>2143</td>
<td>1821</td>
<td>482</td>
<td>74.3</td>
<td>90.4</td>
</tr>
<tr>
<td>4</td>
<td>-7.14</td>
<td>0.997</td>
<td>4801</td>
<td>3361</td>
<td>696</td>
<td>130</td>
<td>90.9</td>
</tr>
</tbody>
</table>

The enrichment of the dataset increased with the number of iterations since sampling at each iteration was performed over the positive dataset derived from the previous iteration, which was obviously a more enriched dataset (Table 5). Thus, with PD2.0 protocol we achieved a final FPDE of 130x while retaining 90% of high-affinity molecules. In the top n enrichment, we evaluated the number of ‘good’ molecules, i.e. with docking scores below the threshold used for
the iteration, within the top n predicted molecules, where the rank was provided by the probability values of the DNN model. Thus, molecules with higher probabilities were expected to be TP with higher confidence, and consequently, top-ranked portions of the dataset would have higher enrichments. As anticipated, enrichment values peaked when considering the top 10 predicted molecules and decrease when considering the top 100, 1000 or the whole predicted dataset. In the final iteration, the corresponding 'top 100' enrichment was 3000x.

Table 6 shows the number of positive samples based on the same threshold value across different iterations. Notably, we observed an exponential increase in positive molecules in the dataset after each iteration, which generated more balanced datasets and allowed us to safely decrease the cutoff value.

Table 6. The total number of molecules of positive class based on a fixed cutoff value in the datasets obtained at different iterations.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>cutoff</th>
<th>Number of Positive molecules (per million)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>71</td>
</tr>
<tr>
<td>2</td>
<td>-7</td>
<td>362</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>1010</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>5752</td>
</tr>
<tr>
<td>End</td>
<td></td>
<td>10370</td>
</tr>
</tbody>
</table>

The distribution of the docking scores after each iteration is plotted in Figure 18. As expected, the overall distribution moved towards molecules with lower docking scores after each
iteration. This also implies that sampling from more enriched datasets after each iteration would give more balanced classes, as shown also in Table 6.

Figure 18. Docking score distribution plot for ER-AF2 at all iterations. The score distribution shifted towards left/better molecules; the shift was more prominent at the initial iteration compared to later ones. This implies that enrichment slowed down as we proceeded through more iterations.

In order to externally validate our model, we docked 10 million molecules randomly sampled from the entire ZINC database. We then used the best model at each iteration to predict their classes (0=bad binders, 1=good binders) and compared these statistics with the internal test set (Table 7). Statistical results were consistent between the internal test set and the external validation set.
After each iteration, we also compared the positive molecules predicted to be TP with the actual TP molecules. We used the precision values of the model to calculate the TP returned by PD2.0 at each iteration and compared it with docking results. Thus, we docked 1 million random molecules from the positive dataset after each iteration, and evaluated the number of molecules below the threshold, as shown in Table 8. The predicted and docking values were consistent, which further reinforced that the model statistics were consistent with the prediction statistics. Unfortunately, in order to assess the consistency of the recall values, we would have to dock the entire ZINC database which would have taken years to complete. For this reason, 10 million compounds were considered as the external validation set.
**Table 8.** A comparison of good molecules (TP) returned by PD2.0 and good molecules returned by docking. The predicted and docking values are very similar.

<table>
<thead>
<tr>
<th>After Iteration</th>
<th>Cutoff</th>
<th>#Good molecules (predicted, per million)</th>
<th>#Good molecules (docking, per million)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-5.24</td>
<td>106000</td>
<td>108000</td>
</tr>
<tr>
<td>2</td>
<td>-5.84</td>
<td>51000</td>
<td>51400</td>
</tr>
<tr>
<td>3</td>
<td>-6.95</td>
<td>7200</td>
<td>70400</td>
</tr>
<tr>
<td>4</td>
<td>-7.14</td>
<td>6400</td>
<td>6000</td>
</tr>
</tbody>
</table>

### 4.3.4.2.2 Conformal prediction

We evaluated the docking results using Conformal Prediction (CP) statistics as described in section 4.2.9. For each iteration, the same model as before is used for CP.

We observed that all the model statistics including ROC-AUC, recall and enrichment values as reported in Table 9 are almost identical to those observed for the threshold-based approximation (Table 5) and the only difference was noticed for the first iteration.

Table 10 compares the model statistics from the internal test set and the external validation set. Almost identical statistics as those of threshold-based approximation were observed again. Based on the above observations we concluded that for our task threshold-based prediction worked on par with conformal prediction.
Table 9. Different model statistics for CP after each iteration for ER-AF2. ROC-AUC, enrichment values and recall values for the best model are reported.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Model cutoff</th>
<th>Top 50</th>
<th>Top 100</th>
<th>Top 1000</th>
<th>FPDE</th>
<th>Model Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-5.24</td>
<td>33</td>
<td>31</td>
<td>27</td>
<td>6.8</td>
<td>74.5</td>
</tr>
<tr>
<td>2</td>
<td>-5.84</td>
<td>168</td>
<td>168</td>
<td>113</td>
<td>11.5</td>
<td>91</td>
</tr>
<tr>
<td>3</td>
<td>-6.95</td>
<td>3428</td>
<td>1821</td>
<td>503</td>
<td>73.2</td>
<td>92.7</td>
</tr>
<tr>
<td>4</td>
<td>-7.14</td>
<td>3637</td>
<td>3839</td>
<td>707</td>
<td>130</td>
<td>90.1</td>
</tr>
</tbody>
</table>

Table 10. Comparing different model statistics for CP between internal test set and external validation set for ER-AF2. ROC-AUC, FPDE, and recall of the best model at each iteration are reported.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Validation Set</th>
<th>Model cutoff</th>
<th>FPDE</th>
<th>Model Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal</td>
<td>-5.24</td>
<td>6.8</td>
<td>74.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6.4</td>
<td>74.6</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>-5.84</td>
<td>11.5</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>12.2</td>
<td>90.1</td>
</tr>
<tr>
<td>3</td>
<td>Internal</td>
<td>-6.95</td>
<td>73.2</td>
<td>92.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>78.3</td>
<td>89.4</td>
</tr>
<tr>
<td>4</td>
<td>Internal</td>
<td>-7.14</td>
<td>130</td>
<td>90.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>136</td>
<td>93</td>
</tr>
</tbody>
</table>

4.3.4.3 Speed analysis

For our ERAF2 study, we used 200-300 CPU cores for docking using GLIDE and 4 GPUs for model making and model predictions. In future applications of PD2.0, one can tweak different
features of the method depending upon the amount of resources available, for example, the number of molecules to be docked at each iteration or the number of hyperparameters to be searched. The overall timeline of the PD2.0 is shown in Table 11. We performed 4 iterations which took about ~12 days and docking of 8 million molecules which were predicted after the last iteration took about 6 days. Thus, the overall time required to virtually screen 1.3 billion molecules against ER-AF2 using PD2.0 was around 18 days. Docking all 1.3 billion molecules using GLIDE with the same amount of resources would have taken about ~2.5 years. Thus, PD2.0 offers a speedup of about 65x. PD2.0 is highly scalable to billions of molecules with expected higher speed and enrichment gains. Figure 19 shows the projection of speedup against a different size of the molecule database.

Table 11. The time required for individual steps of the PD2.0 pipeline. The major time-consuming step is docking.

One PD2.0 iteration takes 3 days using 200-300 CPU cores and 4 GPUs.

<table>
<thead>
<tr>
<th>Steps (per iteration)</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D to 3D optimization</td>
<td>&lt;4 hours</td>
</tr>
<tr>
<td>Docking</td>
<td>2 days</td>
</tr>
<tr>
<td>Model Making</td>
<td>&lt;18 hours</td>
</tr>
<tr>
<td>Prediction</td>
<td>&lt;5 hours</td>
</tr>
<tr>
<td>Overall</td>
<td>~3 days per iteration</td>
</tr>
</tbody>
</table>
Figure 19. Time projection of PD2.0 and regular docking against different database sizes. As the size of the molecule database increases, the difference between the speed of PD2.0 and regular docking becomes more accentuated. Projections were generated with the assumption that when dealing with more molecules, the number of iterations, as well as inference time, will increase. Since the time per iteration is very small due to the low number of molecules required to dock (3 million), increasing by few iterations will still lead to a significant gain in speed compared to docking a library of billions of molecules.

4.3.4.4 PD2.0 on diverse drug targets

We used PD2.0 to screen 570 million molecules from ZINC15 3D against 12 relevant drug targets (Table 1, Chapter 1). The DNN model described previously was trained at each iteration. A basic hyperparameter grid search was performed and the best model was chosen for prediction on the entire dataset (570 million). Docking was performed using OpenEye’s FRED. We generated
an external validation set by docking 9 million molecules randomly sampled from the database against all targets. We then used the best model of each iteration to predict the scores of these 9 million molecules and compared the statistics with the internal test set (test set used during training).

We calculated model ROC-AUC, enrichment values and recall for all the 12 targets at all the iterations (Table 13-24). The enrichment of the dataset with good molecules increased with the iterations, as discussed in the case of ER-AF2 (Figure 35). The results were also consistent between both internal and external validation test sets for all the targets (Table 13-24).

The distribution of docking scores after each iteration is plotted in Figures 23-31. As already observed for ER-AF2, the distributions of scores shifted towards better scoring molecules after each iteration. The mean docking score (calculated on 1 million molecules randomly sampled after each iteration) decreased with the iterations but the rate of decrease slowed down at later iterations. The mean value, as well as the decrease rate, were target-dependent (Figure 33). The number of molecules retained after each iteration is plotted in Figure 34, which shows that the rate of decrease slowed down as well at later iterations, and was also highly target-dependent. The enrichment values (top 10, 100, 1000) were also target-dependent (Figure 35). Model recall, AUC-ROC, and FPDE for the last iteration are reported in Table 12. By using PD2.0 on 12 relevant drug targets, we achieved FPDE values up to ~78X and ’top 100’ enrichment values up to ~2000X.
<table>
<thead>
<tr>
<th>Protein</th>
<th>Model ROC AUC</th>
<th>FPDE</th>
<th>Model Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1T7R</td>
<td>0.97</td>
<td>35</td>
<td>87</td>
</tr>
<tr>
<td>1ERR</td>
<td>0.97</td>
<td>37</td>
<td>90</td>
</tr>
<tr>
<td>5YCP</td>
<td>0.98</td>
<td>38</td>
<td>90</td>
</tr>
<tr>
<td>2ZV2</td>
<td>0.95</td>
<td>18</td>
<td>88</td>
</tr>
<tr>
<td>5L2S</td>
<td>0.98</td>
<td>18</td>
<td>92</td>
</tr>
<tr>
<td>4AG8</td>
<td>0.97</td>
<td>34</td>
<td>88</td>
</tr>
<tr>
<td>5MZJ</td>
<td>0.98</td>
<td>31</td>
<td>90</td>
</tr>
<tr>
<td>6IU</td>
<td>0.97</td>
<td>22</td>
<td>83</td>
</tr>
<tr>
<td>4ZUD</td>
<td>0.97</td>
<td>11</td>
<td>91</td>
</tr>
<tr>
<td>5EKO</td>
<td>0.98</td>
<td>78</td>
<td>90</td>
</tr>
<tr>
<td>4F8H</td>
<td>0.98</td>
<td>48</td>
<td>90</td>
</tr>
<tr>
<td>5OSB</td>
<td>0.98</td>
<td>30</td>
<td>90</td>
</tr>
</tbody>
</table>
Table 13. Model cutoff, ROC-AUC, FPDE and recall for Androgen Receptor (AR) after each iteration. All the statistics are consistent between the internal and external validation set.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Validation Set</th>
<th>Model cutoff</th>
<th>Model ROC AUC</th>
<th>FPDE</th>
<th>Model Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal</td>
<td>-14.00</td>
<td>0.93</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.93</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>-14.74</td>
<td>0.96</td>
<td>9</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.96</td>
<td>9</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>Internal</td>
<td>-15.07</td>
<td>0.96</td>
<td>13</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.96</td>
<td>13</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>Internal</td>
<td>-15.25</td>
<td>0.98</td>
<td>22</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.98</td>
<td>22</td>
<td>88</td>
</tr>
<tr>
<td>5</td>
<td>Internal</td>
<td>-15.63</td>
<td>0.98</td>
<td>35</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.98</td>
<td>35</td>
<td>86</td>
</tr>
</tbody>
</table>

Figure 20. Docking score distribution for AR. The distribution of scores shifted towards better (lower) docking scores after each iteration.
Table 14. Model cutoff, ROC-AUC, FPDE and recall for Estrogen Receptor after each iteration. All the statistics are consistent between the internal and external validation set.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Validation Set</th>
<th>Model cutoff</th>
<th>Model ROC AUC</th>
<th>FPDE</th>
<th>Model Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal</td>
<td>-13.63</td>
<td>0.93</td>
<td>4.3</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.92</td>
<td>4.3</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>-13.88</td>
<td>0.97</td>
<td>9.7</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.96</td>
<td>9.6</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>Internal</td>
<td>-15.00</td>
<td>0.97</td>
<td>37</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.97</td>
<td>34</td>
<td>85</td>
</tr>
</tbody>
</table>

Figure 21. Docking score distribution for ER. The distribution of scores shifted towards better (lower) docking scores after each iteration.
Table 15. Model cutoff, ROC-AUC, FPDE and recall for Peroxisome Proliferator-Activated Receptor (PPARγ) after each iteration. All the statistics are consistent between the internal and external validation set.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Validation Set</th>
<th>Model cutoff</th>
<th>Model ROC AUC</th>
<th>FPDE</th>
<th>Model Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal</td>
<td>-13.00</td>
<td>0.92</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.92</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>-14.04</td>
<td>0.97</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.97</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>Internal</td>
<td>-14.48</td>
<td>0.98</td>
<td>16</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.98</td>
<td>16</td>
<td>91</td>
</tr>
<tr>
<td>4</td>
<td>Internal</td>
<td>-14.70</td>
<td>0.97</td>
<td>23</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.98</td>
<td>23</td>
<td>91</td>
</tr>
<tr>
<td>5</td>
<td>Internal</td>
<td>-15.04</td>
<td>0.98</td>
<td>38</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.98</td>
<td>38</td>
<td>90</td>
</tr>
</tbody>
</table>

Figure 22. Docking score distribution for PPARγ. The distribution of scores shifted towards better (lower) docking scores after each iteration.
Table 16. Model cutoff, ROC-AUC, FPDE and recall for Calcium/calmodulin-dependent protein kinase kinase 2 (CAMKK2) after each iteration. All the statistics are consistent between the internal and external validation set.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Validation Set</th>
<th>Model cutoff</th>
<th>Model ROC AUC</th>
<th>FPDE</th>
<th>Model Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal</td>
<td>-13.70</td>
<td>0.91</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.91</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>-14.47</td>
<td>0.95</td>
<td>7</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.95</td>
<td>7</td>
<td>91</td>
</tr>
<tr>
<td>3</td>
<td>Internal</td>
<td>14.77</td>
<td>0.96</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.97</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>Internal</td>
<td>-14.95</td>
<td>0.96</td>
<td>14</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.97</td>
<td>14</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>Internal</td>
<td>-15.55</td>
<td>0.95</td>
<td>18</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.97</td>
<td>19</td>
<td>90</td>
</tr>
</tbody>
</table>

Figure 23. Docking score distribution for CAMKK2. The distribution of scores shifted towards better (lower) docking scores after each iteration.
Table 17. Model cutoff, ROC-AUC, FPDE and recall for Cyclin-dependent kinase 6 (CDK6) after each iteration.

All the statistics are consistent between the internal and external validation set.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Validation Set</th>
<th>Model cutoff</th>
<th>Model ROC AUC</th>
<th>FPDE</th>
<th>Model Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal</td>
<td>-12.92</td>
<td>0.91</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.91</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>-13.60</td>
<td>0.95</td>
<td>6</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.95</td>
<td>6</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>Internal</td>
<td>-13.87</td>
<td>0.96</td>
<td>8</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.96</td>
<td>8</td>
<td>91</td>
</tr>
<tr>
<td>4</td>
<td>Internal</td>
<td>-13.98</td>
<td>0.97</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.97</td>
<td>10</td>
<td>91</td>
</tr>
<tr>
<td>5</td>
<td>Internal</td>
<td>-14.63</td>
<td>0.98</td>
<td>18</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.98</td>
<td>18</td>
<td>91</td>
</tr>
</tbody>
</table>

Figure 24. Docking score distribution for CDK6. The distribution of scores shifted towards better (lower) docking scores after each iteration.
Table 18. Model cutoff, ROC-AUC, FPDE and recall for Vascular endothelial growth factor receptor 2 (VEGFR2) after each iteration. All the statistics are consistent between the internal and external validation set.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Validation Set</th>
<th>Model cutoff</th>
<th>Model ROC AUC</th>
<th>FPDE</th>
<th>Model Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal</td>
<td>14.13</td>
<td>0.92</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.92</td>
<td>4</td>
<td>89</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>-15.12</td>
<td>0.96</td>
<td>7</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.96</td>
<td>7</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>Internal</td>
<td>-15.41</td>
<td>0.98</td>
<td>16</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.98</td>
<td>16</td>
<td>89</td>
</tr>
<tr>
<td>4</td>
<td>Internal</td>
<td>-15.79</td>
<td>0.97</td>
<td>34</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.96</td>
<td>34</td>
<td>86</td>
</tr>
</tbody>
</table>

Figure 25. Docking score distribution for FEGFR2. The distribution of scores shifted towards better (lower) docking scores after each iteration.
Table 19. Model cutoff, ROC-AUC, FPDE and recall for Adenosine A2a Receptor (ADORA2A) after each iteration. All the statistics are consistent between the internal and external validation set.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Validation Set</th>
<th>Model cutoff</th>
<th>Model ROC AUC</th>
<th>FPDE</th>
<th>Model Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal</td>
<td>12.29</td>
<td>0.91</td>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.91</td>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>-12.99</td>
<td>0.95</td>
<td>8</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.96</td>
<td>8</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>Internal</td>
<td>-13.35</td>
<td>0.95</td>
<td>10</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.95</td>
<td>10</td>
<td>89</td>
</tr>
<tr>
<td>4</td>
<td>Internal</td>
<td>-13.51</td>
<td>0.97</td>
<td>14</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.97</td>
<td>14</td>
<td>91</td>
</tr>
<tr>
<td>5</td>
<td>Internal</td>
<td>-14.07</td>
<td>0.98</td>
<td>31</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.98</td>
<td>31</td>
<td>90</td>
</tr>
</tbody>
</table>

Figure 26. Docking score distribution for ADORA2A. The distribution of scores shifted towards better (lower) docking scores after each iteration.
Table 20. Model cutoff, ROC-AUC, FPDE and recall for Thromboxane A2 Receptor (TBXA2R) after each iteration. All the statistics are consistent between the internal and external validation set.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Validation Set</th>
<th>Model cutoff</th>
<th>Model ROC AUC</th>
<th>FPDE</th>
<th>Model Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal</td>
<td>-15.22</td>
<td>0.89</td>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.89</td>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>-15.83</td>
<td>0.94</td>
<td>5</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.94</td>
<td>5</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>Internal</td>
<td>-16.13</td>
<td>0.96</td>
<td>8</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.96</td>
<td>8</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>Internal</td>
<td>-16.35</td>
<td>0.97</td>
<td>13</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.97</td>
<td>13</td>
<td>89</td>
</tr>
<tr>
<td>5</td>
<td>Internal</td>
<td>-17.00</td>
<td>0.97</td>
<td>22</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.98</td>
<td>23</td>
<td>87</td>
</tr>
</tbody>
</table>

Figure 27. Docking score distribution for TBXA2R. The distribution of scores shifted towards better (lower) docking scores after each iteration.
Table 21. Model cutoff, ROC-AUC, FPDE and recall for Angiotensin II type-1 receptor (AT1R) after each iteration.

All the statistics are consistent between the internal and external validation set.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Validation Set</th>
<th>Model cutoff</th>
<th>Model ROC AUC</th>
<th>FPDE</th>
<th>Model Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal</td>
<td>-13.15</td>
<td>0.90</td>
<td>3</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.89</td>
<td>3</td>
<td>91</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>-13.66</td>
<td>0.93</td>
<td>5</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.93</td>
<td>5</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>Internal</td>
<td>-13.85</td>
<td>0.95</td>
<td>6</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.95</td>
<td>6</td>
<td>89</td>
</tr>
<tr>
<td>4</td>
<td>Internal</td>
<td>-13.97</td>
<td>0.95</td>
<td>7</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.95</td>
<td>7</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>Internal</td>
<td>-14.68</td>
<td>0.97</td>
<td>11</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.97</td>
<td>11</td>
<td>89</td>
</tr>
</tbody>
</table>

Figure 28. Docking score distribution for AT1R. The distribution of scores shifted towards better (lower) docking scores after each iteration.
Table 22. Model cutoff, ROC-AUC, FPDE and recall for Sodium-Ion Channel (Nav1.7) after each iteration. All the statistics are consistent between the internal and external validation set.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Validation Set</th>
<th>Model cutoff</th>
<th>Model ROC AUC</th>
<th>FPDE</th>
<th>Model Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal</td>
<td>-13.93</td>
<td>0.91</td>
<td>3</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.91</td>
<td>3</td>
<td>89</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>-15.32</td>
<td>0.97</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.97</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>Internal</td>
<td>-15.91</td>
<td>0.98</td>
<td>26</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.98</td>
<td>26</td>
<td>89</td>
</tr>
<tr>
<td>4</td>
<td>Internal</td>
<td>-16.43</td>
<td>0.98</td>
<td>78</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.99</td>
<td>78</td>
<td>90</td>
</tr>
</tbody>
</table>

Figure 29. Docking score distribution for Nav1.7. The distribution of scores shifted towards better (lower) docking scores after each iteration.
Table 23. Model cutoff, ROC-AUC, FPDE and recall for bacterial (Gloeobacter) Ligand-gated Ion Channel (GLIC) after each iteration. All the statistics are consistent between the internal and external validation set.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Validation Set</th>
<th>Model cutoff</th>
<th>Model ROC AUC</th>
<th>FPDE</th>
<th>Model Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal</td>
<td>-10.04</td>
<td>0.93</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.92</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>-10.82</td>
<td>0.97</td>
<td>11</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.97</td>
<td>11</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>Internal</td>
<td>-11.27</td>
<td>0.98</td>
<td>23</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.98</td>
<td>23</td>
<td>91</td>
</tr>
<tr>
<td>4</td>
<td>Internal</td>
<td>-11.58</td>
<td>0.98</td>
<td>48</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.98</td>
<td>48</td>
<td>90</td>
</tr>
</tbody>
</table>

**Figure 30.** Docking score distribution for GLIC. The distribution of scores shifted towards better (lower) docking scores after each iteration.
Table 24. Model cutoff, ROC-AUC, FPDE and recall for Gamma-AminoButyric Acid receptor subunit alpha-1 (GABA α1) after each iteration. All the statistics are consistent between the internal and external validation set.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Validation Set</th>
<th>Model cutoff</th>
<th>Model ROC AUC</th>
<th>FPDE</th>
<th>Model Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal</td>
<td>-9.74</td>
<td>0.90</td>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.90</td>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>-10.26</td>
<td>0.95</td>
<td>6</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.95</td>
<td>6</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>Internal</td>
<td>-10.53</td>
<td>0.96</td>
<td>9</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.96</td>
<td>9</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>Internal</td>
<td>-10.68</td>
<td>0.96</td>
<td>12</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.96</td>
<td>12</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>Internal</td>
<td>-11.23</td>
<td>0.98</td>
<td>30</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td></td>
<td>0.97</td>
<td>29</td>
<td>87</td>
</tr>
</tbody>
</table>

Figure 31. Docking score distribution for GABA α1. The distribution of scores shifted towards better (lower) docking scores after each iteration.
Figure 32. Top 10, 100, 1000 enrichment for the 12 targets after the last iteration. The enrichment values decreased for an increasing number of top molecules, showing that the DNN model was able to prioritize high-affine molecules (good) i.e. molecules associated to higher DNN probabilities had higher chances of being good-scoring molecules.
**Figure 33.** The mean docking score distribution for the 12 targets across all the iterations. The mean score decreased with the iterations. The decrease rates were higher at the beginning and eventually slowed down. The rate of decrease, as well as the docking score values, were protein-dependent.
Figure 34. The number of molecules left after each iteration for the 12 targets. The number of molecules left after each iteration decreased with the iterations. The decrease was more significant at the beginning and eventually slowed down, and it was highly protein-dependent.
**Figure 35.** Top 100 enrichment value for the 12 targets after each iteration. The enrichment value increased with the iterations and the rate of increase, as well as maximum enrichment value, was protein-dependent.

### 4.4 Future direction

PD2.0 is based on multiple stages: 2D to 3D conversion, molecular docking, model making, hyperparameter optimization, and prediction. Out of all these components, 2D to 3D conversion and molecular docking have little room for improvement. In our opinion, the type of model and the hyperparameter search are the two areas that can further be improved to increase the speedup and accuracy of PD2.0. For example, graphical neural networks (GNNs) implemented into the deepchem framework could replace the feed-forward NNs, where SMILES can be represented as graphs, and other methods (Bayesian optimization, gradient-based optimization, genetic optimization or population-based optimization methods) could be used instead of grid-
based hyperparameter search to select a better set of hyperparameters and provide a more accurate model.

PD2.0 is protein-specific i.e. it needs to be started from the beginning for every new target and we use only the information of the molecule. In the future, we would want to include information about the protein as well as the available experimental information. This can help us to test the molecules against multiple proteins from the same family and improving the specificity. One of the major limitations of PD2.0 is that the rate of decrease of the molecule database reduces significantly after the first iteration. A better way of sampling (i.e. sampling more molecules similar to the incorrect predictions) may help in tackling this problem.
Chapter 5: Virtual screening of ETS transcription factor ERG with PD2.0

5.1 Introduction

In this chapter, we have discussed ETS transcription factor ERG and how it is linked to prostate cancer, then we discussed the small molecule inhibitor that we found previously and finally, we applied our developed method PD2.0 to virtually screen a large-scale 3D molecule database for identifying new classes of small-molecule inhibitor.

The common driver of prostate cancer (PCa) is androgen signaling and is often treated by Androgen Receptor (AR) pathway inhibitors. The treatment leads to a 30% to 40% decrease in disease-specific mortality but almost all the patients end up developing resistance against the AR pathway inhibitors, demanding the development of novel therapeutics.

ERG is a transcription factor with a full-length of 486 amino-acid and a molecular weight of 54kDa. ERG is a member of the ETS family, the majority of which have the ETS DNA-binding domain (DBD) that is around 85 amino acid long. The ETS domain recognizes the DNA sequence containing a core GGA(A/T) motif. ERG is initially highly expressed in the embryonic mesoderm and endothelium where it plays an important role in bone development, the formation of urogenital tract and vascular system. ERG is also expressed at high levels in embryonic neural crest cells during their migratory phase. ERG expression decreases during vascular development but continues to regulate the pluripotency of hematopoietic stem cells, endothelial cell (EC) homeostasis and angiogenesis.

The most prevalent genetic irregularity occurring in ~50% of PCa patients is the fusion of transmembrane protease serine 2 (TMPRSS2)-ERG gene. ERG is not expressed by prostatic epithelial cells but the fusion of the open reading frame sequence of ERG with the AR response element of TMPRSS2 results in a 6000-fold increase (after androgen stimulation) in the expression
level of ERG, making it one of the most commonly overexpressed genes in PCa\textsuperscript{150}. This overexpression of ERG in combination with loss of tumor suppressor such as PTEN leads to tumor tumorigenesis in both xenograft and transgenic mouse models\textsuperscript{151}. ERG modulates various PCa related phenotypes such as disruption of the epithelial differentiation program via AR dysregulation\textsuperscript{109}, activation of c-Myc, epigenetic reprogramming via EZH2\textsuperscript{110} and promotion of genomic instability via PARP dysregulation\textsuperscript{111}. Overexpression of ERG also promotes epithelial-mesenchymal transition (EMT) and enables the transformed cells to acquire migratory and invasive characteristics\textsuperscript{109, 112}. Androgen receptor is responsible for the initiation of ERG expression initially but later on, ERG can become self-driven through feed-forward regulation\textsuperscript{152}. In summary, there are multiple lines of evidence that together strongly demonstrate that ERG is an attractive molecular target for developing PCa therapies. A more detailed review can be found in the paper by Hsing et.al\textsuperscript{153}.

We have previously\textsuperscript{79} discovered and characterized the new class of ERG small molecule inhibitors using CADD. The starting database size for virtual screening was about 3 million molecules (extracted from the ZINC database in 2013), out of which 48 compounds were screened \textit{in vitro} that led to the discovery of compound VPC-18005. VPC-18005 inhibited ERG-driven transcriptional activity in VCaP and PNT1B-ERG cells, and the interaction between ERG and VPC-18005 was also validated using NMR spectroscopy (consistent with the proposed in-silico model). Similar to the TOX project, we have previously only used about 1% of available molecules for virtual screening against ERG. Thus, expanding virtual screening from the initial 3 million to 100s of millions of molecules provides a tremendous opportunity for ERG drug discovery. We have now applied our developed method PD2.0 to virtually screen more than 300 million molecules (all the molecules with 3D representation at the time of \textit{in silico} screening in 2018-
2019) against ERG to select the top 5 million candidate molecules for subsequent computational steps such as consensus scoring.

5.2 Methods

We used PD2.0 developed as described in Chapter 4 to screen the non-conventional target ERG. We used all the 3D conformations available in the ZINC15 database at the time of screening, which accounted for 350 million molecules. The ERG protein (in complex with VPC-18005) was prepared and optimized using Protein Preparation Wizard of Schrödinger suite. The small molecule binding pocket was defined based on the same docking pose of previously identified hit compound VPC-18005, using the Schrödinger’s Grid Preparation Wizard. Docking was performed using Schrödinger’s GLIDE program (Standard Precision mode with default parameters).

5.3 Results and discussion

We used PD2.0 to screen the non-conventional target, ERG which is a DNA binding protein (transcription factor). However, transcription factors have always been very difficult to target because they exert their activity via protein-protein and/or protein-DNA interactions. Unlike proteins such as enzymes, which have a well-defined ligand-binding site that small molecules can target, the extensive surface-exposed sites of transcription factor interactions are usually difficult to disrupt by using VS. We have a good track record of successfully identifying SMIs inhibiting protein-DNA interactions developed via CADD to target the DNA-binding domains of other cancer drug targets such as AR, MYC, and TOX. To target ERG we used PD2.0 that we developed in the previous section.
For ERG, we performed 3 iterations of Progressive Docking, and Figure 36 shows the docking score distribution after each iteration. The plot shows that the distribution moves towards the left with lower (better) docking scores, indicating the molecule dataset is more enriched with good molecules after each iteration. The average docking score improves from -4 to -5.8, and the rate of improvement was higher initially and gradually slows down as Progressive Docking approached towards the final iteration. Table 25 shows the model statistics for ERG after each iteration, and the model enrichment (FPDE) improves after each iteration. We performed 3 iterations of PD2.0 that took about 9 days, and a total of 5 million molecules remained after the final iteration. The docking of this final set of 5 million molecules, which has the best binding potential against ERG as determined by PD2.0, took another 3 days. Thus, the total amount of time required to virtually screen 350 million molecules against ERG was only 12 days by using PD2.0, and it would have taken more than 300 days by using a regular docking method without any form of machine learning. The application of PD2.0 on large-scale virtual screening against ERG has provided a 25x speed boost and an FPDE of 72x.

When we compared the docking scores of the top molecules obtained from PD2.0 and the top molecules obtained from just standard molecular docking, there are 100 times more molecules with docking scores less than -6.5 using PD2.0 compared to regular docking. The docking score is on a Log10 scale, and each score unit difference represents a 10-fold difference in binding affinity. Thus, the best molecule (docking score of -8.9) obtained using PD2.0 is theoretically 100 times better than the best molecule obtained using regular docking (docking score of -7). In addition, the average docking score of the top 5 million molecules from PD2.0 (mean docking score of -5.8) is 100 times better than the mean docking score of 3 million molecules from regular docking (mean docking score of -3.8, from initial screening). Based on the docking score values,
we anticipate that PD2.0 can provide more potent ERG inhibitors compared to VPC-18005, which has a docking score around -6.

In addition to being a critical drug target in prostate cancer, ERG has been shown to play an important role in other types of cancers including Ewing sarcoma and Leukemia\textsuperscript{154-156}. Ewing (or Ewing's) sarcoma (EWS) is the second most common bone cancer in children and young adults, and it is associated with high metastatic potential\textsuperscript{154}. Leukemia represents a group of cancers that begin in the bone marrow and result in high numbers of abnormal white blood cells. Thus, the development of better ERG SMIs through the use of PD2.0 will not only enable better therapeutics for prostate cancer but also provide new therapies to target many other cancer types including Ewing sarcoma and Leukemia.

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Docking score cutoff</th>
<th>Model ROC AUC</th>
<th>FPDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-6.0</td>
<td>0.81</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>-6.5</td>
<td>0.90</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>-6.65</td>
<td>0.90</td>
<td>72</td>
</tr>
</tbody>
</table>

Table 25. Model statistics after each iteration for ERG. ROC-AUC and FPDE values for the best model are reported. We observed an increase in all enrichment values after each iteration. The model cutoff/definition of good molecules improved after each iteration as well.
Figure 36. Docking score distribution for ERG. The distribution of scores shifted towards better (lower) docking scores after each iteration. Lower docking score is an indication of a molecule with better binding affinity with ERG.

5.4 Future directions

Through the use of PD2.0, we have virtually screened 350 million molecules against ERG and identified the top 5 million molecules based on the GLIDE docking scores. In the future, we will perform consensus scoring with additional docking programs by using the in-house CADDpipeline (Chapter 1) to further narrow this set of molecules down to the best candidates (around 100 compounds) for experimental testing. In particular, all of the 5 million molecules will be further docked with the FRED program from OEDocking (up to 500 conformers will be
generated for each molecule and will be docked using FRED with default parameters), and the corresponding RMSD values will be calculated for the top poses. All the molecules with an RMSD ≤ 2Å will be retained and will be docked again using another docking program, ICM. For the poses predicted by ICM (default parameters), RMSD values against Glide will be calculated and only the molecules with RMSD ≤ 2Å will be retained.

Within the resulting set, theoretical pKi values will be calculated for each molecule using a custom MOE SVL script. Other properties like ADMET (absorption, distribution, metabolism, excretion, toxicity) and pharmacokinetics predictions will also be calculated by using computational programs such as ADMET Predictor, FAF-Drugs and Quantitative Estimate of Drug-likeness (QED).

As the next step, a consensus scoring method will be used, and molecules with total vote greater than a threshold value will be retained and then clustered together to remove the similar compounds (70% similarity). Finally, around 100 compounds will be selected for in vitro testing. We anticipate this will lead to new classes of anti-ERG small molecules, with better activities than compound VPC-18005.
Chapter 6: Conclusions

While drug discovery continues to be an expensive and time-consuming effort, the thesis work shows this process can be significantly accelerated through the combined use of molecular docking and machine learning, as demonstrated on two important cancer drug targets, TOX and ERG.

TOX protein represents a promising drug target for CTCL for which there is still no small-molecule inhibitor developed. To address this challenge, we employed the CADD technology and virtually screened 7.6 million molecules, firing them into the DNA binding domain of TOX. As a result, we identified and experimentally confirmed 18 compounds with micro-molar potency and sufficient selectivity toward TOX. The success of the initial campaign highlights the need to screen a much larger available library of virtual molecules (approximately 1 billion entries) for identification of additional molecules with improved anti-TOX potentials. Unfortunately, docking 1 billion compounds into TOX target site would take years of processing time on even advanced computer cluster available at the VPC.

To address this challenge, we developed a novel AI-enabled CADD method called PD2.0, which allows screening billions of molecules using a very reasonable amount of time and computational resources.

Thus, within ~18 days we were able to screen a library of 1.3 billion molecules against ER-AF2 target using 300 CPU and 4 GPUs cores, achieving a VS enrichment of >100 and a speedup of 65 folds. We also demonstrated that PD2.0 can be applied to a wide variety of drug targets and docking protocols. For instance, we successfully used the developed approach to dock 570 million molecules against 12 targets from 4 main target protein (3 for each), including G protein-coupled receptors, kinases, ion channels, and nuclear receptors.
Furthermore, we have applied PD2.0 to ERG, an important drug target not only in prostate cancer but also in Ewing sarcoma and Leukemia. Through the use of PD2.0, we have virtually screened 350 million molecules against ERG and identified the top 5 million molecules based on the GLIDE docking scores. We anticipate applying subsequent CADD protocols will narrow these 5 million molecules down to 100 best drug candidates with potent anti-ERG activities.

To summarize, we have developed PD2.0 as a highly-scalable method enabling much higher speed and enrichment gains for virtual screening. Due to the exponential growth of databases such as ZINC and Enamine REAL, we expect PD2.0 to play a pivotal role in upcoming challenges faced by CADD community. We will continue working on making PD2.0 more accurate and scalable.
Bibliography


33. ENAMINE. https://enamine.net/library-synthesis/real-compounds/real-database.


58. Hechenbichler K, Schliep K. Weighted k-nearest-neighbor techniques and ordinal classification. 2004


70. Cherkasov A, Muratov EN, Fourches D, et al. QSAR modeling: where have you been? Where are you going to? *Journal of medicinal chemistry*. 2014;57(12):4977-5010
77. ADMET Predictor [computer program]. Simulations Plus; 2018.

103


