Computational ligand discovery for the human and zebrafish sex hormone binding globulin

by

Nels Thorsteinson

B.Sc., Queen's University, Kingston ON, 2005

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

MASTER OF SCIENCE

in

THE COLLEGE FOR INTERDISCIPLINARY STUDIES

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

June 2008

ABSTRACT

Virtual screening is a fast, low cost method to identify potential small molecule therapeutics from large chemical databases for the vast amount of target proteins emerging from the life sciences and bioinformatics. In this work, we applied several conventional and newly developed virtual screening approaches to identify novel non-steroidal ligands for the human and zebrafish sex hormone binding globulin (SHBG).

The 'benchmark set of steroids' is a set of steroids with known affinities for human SHBG that has been widely used for validation in the development of different virtual screening methods. We have updated this data set by including additional steroidal SHBG ligands and by modifying the predicted binding orientations of several benchmark steroids in the SHBG binding site based on the use of an improved docking protocol and information from recent crystallographic data. The new steroid binding orientations and the expanded version of the benchmark set was then used to create new in silico models which were applied in virtual screening to identify high-affinity non-steroidal human SHBG ligands from a large chemical database.

Anthropogenic compounds with the capacity to interact with the steroid-binding site of SHBG pose health risks to humans and other vertebrates including fish. We constructed a homology model of SHBG from zebrafish and applied virtual screening to identify ligands for zebrafish SHBG from a set of 80 000 existing commercial substances, many of which can be exposed to the aquatic environment. Six hits from this in silico screen were tested experimentally for zebrafish SHBG binding and three of them, hexestrol, 4-tert-octylcatechol, dihydrobenzo(a)pyren-7(8H)-one demonstrated micromolar binding affinity for the zebrafish SHBG.

These findings demonstrate the feasibility of using virtual screening to identify anthropogenic compounds that may disrupt or highjack functionally important protein:ligand interactions. Studies applying this new computational toxicology method could increase the awareness of hazards posed by existing commercial chemicals at relatively low cost.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Many thanks to Dr. Geoffrey L. Hammond, Dr. Johannes Müllegger, Dr. Fuqiang Ban, and to Dr. Osvaldo Santos-Filho for their teachings in the fields related to this work.

I am very grateful for my supervisor, Dr. Artem Cherkasov, whose teachings and guidance has accelerated my development leading to many achievements including this work. With financial support from the CIHR bioinformatics training program, I was able to spend a summer working with Dr. Cherkasov's colleague, Dr. Mikhail Gelfand, in Moscow, Russia. This work and travel experience turned out to be invaluable for my personal and professional development.

DEDICATION

To my parents and grandparents

CO-AUTHORSHIP STATEMENT

The manuscript in chapter 2 is a version of a work published in The Journal of Medicinal Chemistry co-authored by Artem Cherkasov, Fuqiang Ban, Osvaldo Santos-Filho, Nels Thorsteinson, Magid Fallahi, and Geoffrey L. Hammond. This research was a collaborative effort whereby the computational discoveries and generation of figures was conducted by Artem Cherkasov, Osvaldo Santos-Filho, Fuqiang Ban, and Nels Thorsteinson. Magid Fallahi, under the supervision of Geoffrey L. Hammond, performed the experimental validation in the form of a human sex hormone binding globulin (SHBG) ligand binding assay. The research was conceived by Artem Cherkasov, who also prepared the original manuscript. Nels Thorsteinson made major edits to Artem Cherkasov's manuscript and included it as chapter 2 of this work.

The manuscript included in chapter 3 is a version of a work submitted to Toxicology and Applied Pharmacology that was co-conceived by Artem Cherkasov and Geoffrey L. Hammond and was co-authored by Nels Thorsteinson, Fuqiang Ban, Osvaldo Santos-Filho, Solange Miguel-Queralt, Caroline Underhill, Artem Cherkasov, and Geoffrey L. Hammond. This research was a collaborative effort whereby the computational discoveries and figure generation was conducted by Nels Thorsteinson with the exception of constructing the CoMFA and CoMSIA models which was done by Fuqiang Ban and Osvaldo Santos-Filho. Solange Miguel-Queralt and Caroline Underhill, under the supervision of Geoffrey L. Hammond, performed the experimental validation in the form of a zebrafish SHBG ligand binding assays. The research on the occurrence and uses of the identified anthropogenic zebrafish SHBG ligands was performed by Seyed M.H. Tabaei. The manuscript was prepared by Nels Thorsteinson and was edited by Artem Cherkasov and then extensively by Geoffrey L. Hammond.

1 Introduction

1.1 Life Sciences, bioinformatics, and computer-aided drug design

Recently developed high-throughput technologies for the life sciences such as rapid DNA sequencing, genotyping arrays, microarrays, and mass spectrometry continue to deliver massive amounts of data. Bioinformatitions have been applying a number of computational methods to these data such as data mining or protein network modeling in order to identify potential drug target proteins.

Finding therapeutics to act on potential drug targets is a challenging and often very expensive process. Moreover, much of the research in this area is funded by and confined within the pharmaceutical industry. Pharmaceutical companies typically aim to deliver treatments for which the monetary market is large, and not to investigate lower market treatments, such as treatments for diseases plaguing developing countries. It is important to allocate more resources and to improve affordable methods for the academic discovery of treatments for such diseases.

In recent years, computer-aided drug design, which utilizes computer models and computational chemistry for processes such as virtual screening, has been increasing in popularity in academia. This field offers a low cost method to identify therapeutics for the vast amount of target proteins emerging from life science and bioinformatics laboratories.

1.2 Computer-aided drug design

One of the most critical parts of computer-aided drug design is the theoretical identification of small molecules that can bind to a given target biomolecule and thus produce a desired therapeutic effect. Typically this requires performing fast

computational chemistry calculations on an experimentally defined structure or a computer model of the target and on a set of potential ligands. Often, large chemical databases are screened in silico to dramatically reduce the databases to small subsets of compounds that are highly likely to bind to the target. These smaller sets are to be synthesized and tested for their effect on the target experimentally and then lead optimization strategies can be applied to the new known ligands to produce structures with an even higher affinity for the target, or, if no sufficient leads are produced by the initial screen, then the new known ligands can be used to build better models that can be used in a more focused screen that would be more likely to produce sufficient leads. A basic schematic overview of this general process is shown in Figure 1.1.

Structures of protein targets are usually obtained by established methods such as X-ray crystallography or NMR, however, when no structure is available for a given receptor, structure prediction can often be applied to produce an adequate model. If the structure of a homologous protein is available, homology modeling, the most accurate form of protein structure prediction, can be applied to accurately predict a protein's structure from that of a homologue. For this, the sequences of the proteins are aligned, and the residues of the template structure of the homologous protein are replaced by those from the sequence of the protein being modeled. The resulting model is often refined by energy minimization in order that its atoms rest in the conformation of a potential energy minima (Flohil et al., 2002).

Energy minimization in molecular modeling is performed in order to position the atoms in a low potential energy state. Energy minimization consists of finding a set of atomic coordinates that correspond to a local minimum of a molecular energy function. This is done by applying optimization techniques such as steepest descent or conjugate

gradient to minimize the forcefield potential energy function. A forcefield is the function and parameter sets used to describe the potential energy of a system. Several different forcefields have been developed and continue to be developed in efforts to produce more accurate calculations (Pearlman et al., 1981; Brooks et al., 1983; Halgren, 1996; Jorgensen and Tirado-Rives, 1988).

Homology models are often further refined by molecular dynamics (MD) simulations which incorporate forcefield calculations and Newton's equations of motion to simulate the atoms of a system moving in time. For this, water molecules are added to the protein and this system is first energy minimized. Then the MD simulation is carried out by calculating the forces being applied to each atom and calculating their displacements and then moving each atom over the next time step. The duration of time steps vary but are usually between 1 ps and 2 ps. This process is repeated several times until a simulation time typically greater than 5 ns is complete. If a stable MD structure is reached, then this model is used for further computational methods. MD is also used in the context of computational drug design to assess the flexibility of a target protein and its ability for induced fit ligand binding (Koshland, 1994; Sotriffer et al., 2004).

Molecular docking is the process of predicting the binding orientation and affinities of compounds and is commonly used in virtual screening. Various academic and commercial docking packages are available for virtual compound screening such as the Glide docking program (Schrödinger, Inc., 2006), which consistently ranks among the best docking programs (Perola et al., 2004; Kellenberger et al., 2004; Warren et al., 2006; McGaughey et al., 2007). Glide works by performing a search of many ligand orientations and conformations in the receptor's active site, called poses. Each pose is

given a score by use of a scoring function that incorporates Coulomb and Lennard-Jones interactions, hydrogen bonding, and hydrophobic interactions. The higher scoring poses are refined by energy minimization and scored again. In the end, the highest scoring pose for each ligand along with its Glide score is returned (Friesner et al., 2004; Halgren et al., 2004).

Another important tool for virtual screening is quantitative structure activity relationship (QSAR) techniques that work by measuring a set of structural descriptors for a set of compounds with known affinities for the target (often called the training set) then using optimization algorithms to derive an equation where affinity is a function of these descriptors. This equation is then used to predict the activity of an external set of test compounds. Two such broadly used techniques called comparative molecular field analysis (CoMFA) (Cramer et al., 1988) and comparative molecular similarity index analysis (CoMSIA) (Klebe et al., 1994) were originally validated using the 'steroid benchmark set' of sex hormone binding globulin (SHBG) ligands.

The CoMFA method works by calculating the steric and electrostatic fields for each compound at each point in a grid surrounding the set of compounds. Using mathematical methods such as partial least squares, these fields are used to arrive at a function for estimating the binding constant. This normally results in a number of the fields being removed and only a few key fields end up being important for estimating binding affinity. CoMSIA works in a similar fashion but includes hydrophobic and hydrogen bond donor and acceptor fields in addition to steric and electrostatic fields. An illustrated example of a CoMSIA model built from the benchmark set of steroids docked to the SHBG binding site of the 1LHN entry (Grishkovskaya et al., 2002a) of the Protein Data Bank (PDB) (Berman, et al., 2000) is shown in Figure 1.2. Cross-validation is often

used to validate that the CoMFA/CoMSIA models are adequate for predicting the binding constants for external test ligands. For cross-validation, typically 90 % of the training set compounds will be used to build the model, and the accuracy of the predictions for the remaining 10 % is obtained. This process is repeated for several different divisions of the training set and an estimate of the overall accuracy of the models is obtained.

Another example of QSAR descriptors successfully used for describing SHBG binding compounds is the set of parameters called 'inductive' descriptors (Cherkasov et al., 2005b; Cherkasov, 2005) which quantify the inductive effect of the electronegative atoms in a molecule. The inductive effect is the effect of the transmission of charge through a chain of atoms in a molecule by electrostatic induction. Inductive QSAR descriptors that have been derived from free energy equations for inductive and steric substituents have been used successfully in virtual screening (Cherkasov et al., 2005a).

The affinity of a ligand for a receptor can be estimated by calculating the binding free energy (ΔG_{bind}) of a ligand-receptor complex using, for instance, the linear interaction energy (LIE) method (Hansson et al., 1988). For the LIE method, two MD simulations are carried out, one with the ligand free in solution and one where it is bound to the solvated receptor. From these simulations, the change in free energy resulting from the ligand binding to the receptor is calculated by the LIE formula:

$$\Delta G_{bind} = \alpha\Delta\langle V_{vdw}\rangle + \beta\Delta\langle V_{el}\rangle + \gamma$$

$\Delta\langle V_{vdw}\rangle$ and $\Delta\langle V_{el}\rangle$ are the difference between the bound and free states of the MD averages of the ligand-surrounding van der Waals and electrostatic interaction energies, respectively. The α, β, and γ parameters are used to scale the van der Waals energies, electostatic energies and hydrophobicity respectively.

1.3 The sex hormone binding globulin

SHBG is the major sex steroid transporter in the plasma of a wide range of species (Westphal, 1986). It is produced by hepatocytes and it regulates the access of sex steroids to their target tissues (Siiteri et al., 1982; Hammond, 1995). SHBG genes are expressed in the gut and liver of zebrafish (*Danio rerio)* embryos and in the intestine until the fish mature (Miguel-Queralt et al., 2004). Small amounts of SHBG mRNA have been detected in zebrafish testes (Miguel-Queralt et al., 2004) where it is likely to influence spermatogenesis (Joseph, 1994). In sea bass, SHBG is expressed very early in development and plasma SHBG concentrations vary according to reproductive seasons (Miguel-Queralt et al., 2007). All of this suggests that SHBG influences proper reproduction and development in these species.

It has also been found that numerous human diseases such as endometrial cancer, ovarian dysfunction, male and female infertility, osteoporosis, diabetes, and cardiovascular diseases are associated with abnormal SHBG levels, implicating the protein as a prospective drug target (Nisker et al., 1980; Hogeveen et al., 2002; Anderson, 1974; Van Pttelgergh et al., 2004; Rapuri et al., 2004; Haffner et al., 1989).

The ligand binding characteristics of the human SHBG (hSHBG) have been extensively studied by X-ray crystallography (Grishkovskaya et al., 2002a; Grishkovskaya et al., 2002b), and by mutagenesis (Hammond et al., 2003; Avvakumov et al., 2002). A high level of variation has been observed amongst the binding constants of the steroid benchmark set of hSHBG ligands.

The steroid binding site is located in the laminin G4-like (LG-4) domain of hSHBG and X-ray crystallography structures of this domain and a number of different steroidal ligands bound to the binding site have been published. The highest resolution hSHBG

crystal structure (1KDM, resolved at 2.35 Å [Grishkovskaya et al., 2002b]) available from PDB contains dihydrotestosterone (DHT) bound to the active site.

1.4 Motivation and research goals

Formerly, the original CoMFA and CoMSIA models developed using the steroid benchmark set were done so assuming that the steroid scaffolds of the training set were all aligned in the same orientation in hSHBG, however, recent crystallography data indicates that estradiol derivatives bind in the opposite orientation to androgens (Grishkovskaya et al., 2002a). We used this information to form a newly-aligned steroid benchmark set and used it to redevelop and evaluate the impact of the steroid realignment of the training compounds on the performance of CoMFA and CoMSIA models.

Diseases caused by a limited availability of a sex steroid, such as osteoporosis, can possibly be treated by high-affinity non-steroidal hSHBG ligands, provided that they can displace steroids, such as estradiol, and increase their availability as needed by the tissues. The new CoMFA and CoMSIA models and QSAR models based on inductive descriptors as well as Glide docking were used to screen the large ZINC chemical database for potential non-steroidal hSHBG ligands. A number of the hits from this screen were experimentally tested for hSHBG binding, some of which were found to have a relatively high affinity for hSHBG.

Growing concerns about the impact of anthropogenic compounds on the environment has prompted more research into the potential toxicity of commercial chemicals. The uptake of steroids by zebrafish from their aquatic environment has been reported to be influenced in some way by their affinity for SHBG (Scott et al., 2005),

therefore, any zebrafish SHBG (zfSHBG) binders in the water pose a risk to the fish. We applied similar virtual screening techiques to those used for hSHBG to a zfSHBG homology model and identified micromolar zfSHBG ligands from a database of existing commercial substances.

It has previously been found that the affinities of steroids for hSHBG and zfSHBG vary considerably (Miguel-Queralt et al., 2004). We apply MD and binding free energy calculations on the zfSHBG model and on several mutant models to explain the observed ligand binding disparities between hSHBG and zfSHBG.

Figure 1.1

A basic overview of computer-aided drug design

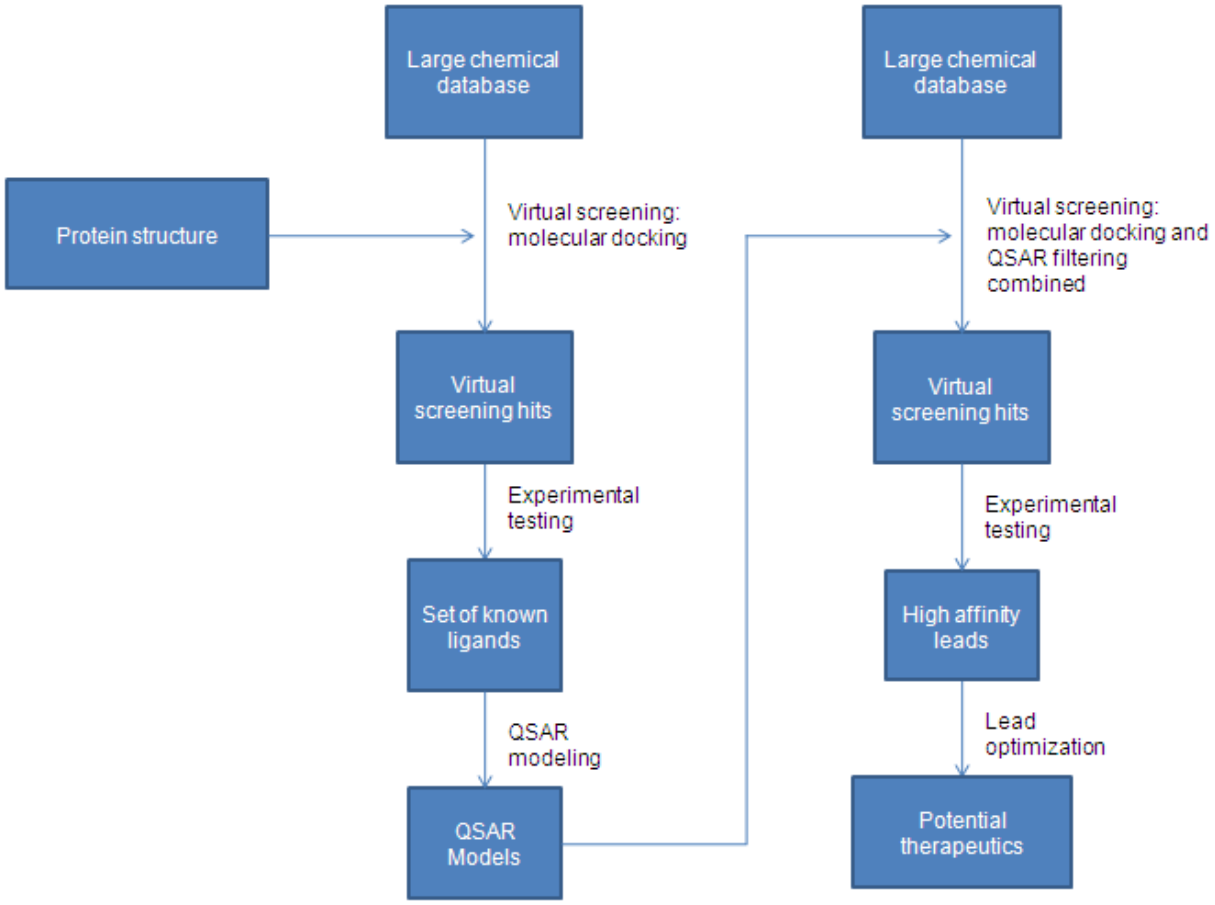
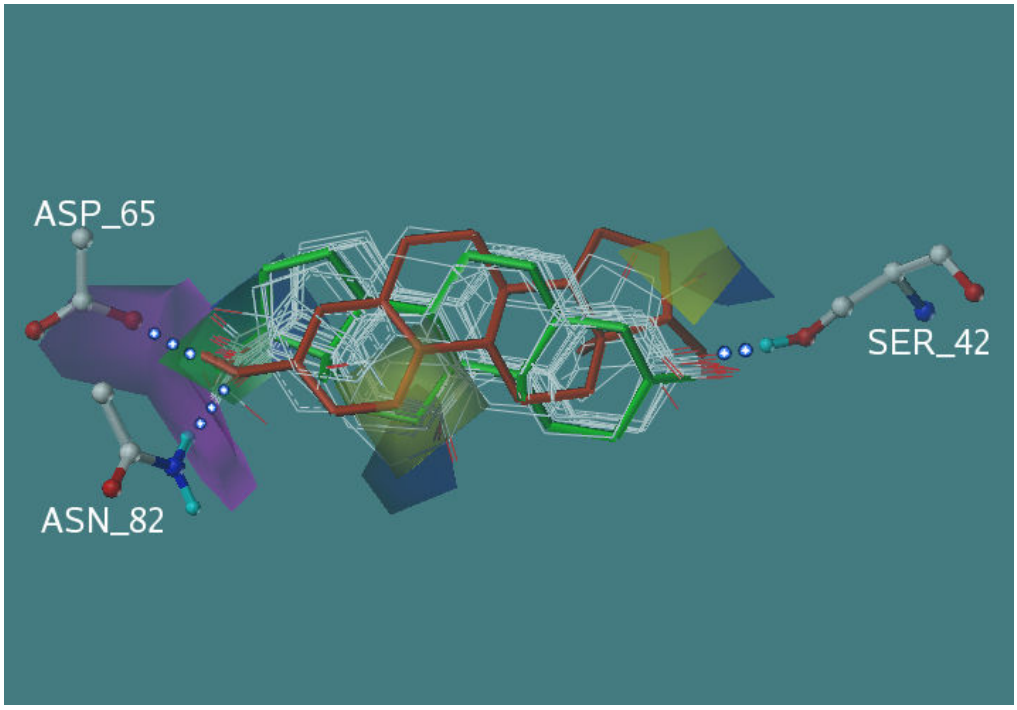


Figure 1.2

The CoMSIA fields derived from the benchmark steroids.



The fields are superimposed with the docked structure of estrogen (red) and DHT (green) inside the hSHBG active site (derived from 1LHU and 1LHN crystal structures respectively [Grishkovskaya et al., 2002a]). The Glide docking poses of the benchmark steroids (white) are also shown.

1.5 References

Anderson, D.C. (1974). Sex-hormone-binding globulin. *Clin. Endocrinol*. **3**, 69-96.

Avvakumov, G. V., Grishkovskaya, I., Muller, Y. A., and Hammond, G. L. (2002). Crystal structure of human sex hormone-binding globulin in complex with 2-methoxyestradiol reveals the molecular basis for high affinity interactions with C-2 derivatives of estradiol. *J. Biol. Chem.* **277**, 45219-45225.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217

Cherkasov, A., Shi, Z., Fallahi, M., and Hammond, G. L. (2005a). Successful in silico discovery of novel nonsteroidal ligands for human sex hormone binding globulin. *J. Med. Chem.* **48**, 3203-3213.

Cherkasov, A., Shi, Z., Li, Y., Jones, S. J., Fallahi, M., and Hammond, G. L. (2005b). 'Inductive' charges on atoms in proteins: comparative docking with the extended steroid benchmark set and discovery of a novel SHBG ligand. *J. Chem. Inf. Model* **45**, 1842-1853.

Cherkasov A. (2005). 'Inductive' Descriptors. 10 Successful Years in QSAR. *Curt. Comp-Aided Drug Design* **1**, 21-42.

Cramer, R. D., Patterson, D. E., and Bunce, J. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110**, 5959-5967.

Flohil, J. A., Vriend, G., and Berendsen, H. J. (2002). Completion and refinement of 3-D homology models with restricted molecular dynamics: application to targets 47, 58, and 111 in the CASP modeling competition and posterior analysis. *Proteins* **48**, 593-604.

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47**, 1739-1749.

Grishkovskaya, I., Avvakumov, G. V., Hammond, G. L., Catalano, M. G., and Muller, Y. A. (2002a). Steroid ligands bind human sex hormone-binding globulin in specific orientations and produce distinct changes in protein conformation. *J. Biol. Chem.* **277**, 32086-32093.

Grishkovskaya, I., Avvakumov, G. V., Hammond, G. L., and Muller, Y. A. (2002b).

Resolution of a disordered region at the entrance of the human sex hormone-binding globulin steroid-binding site. *J. Mol. Biol.* **318**, 621-626.

Haffner, S. M., Katz, M. S., Stern, M. P., and Dunn, J. F. (1989). Association of decreased sex hormone binding globulin and cardiovascular risk factors. *Arteriosclerosis* **9**, 136-143.

Halgren, T.A. (1996). The Merck Force Field. *J. Comp. Chem.* **17** 490-512, 520-552, 553-586, 587-615, 616-641.

Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., and Banks, J. L. (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **47**, 1750-1759.

Hammond, G. L., Avvakumov, G. V., and Muller, Y. A. (2003). Structure/function analyses of human sex hormone-binding globulin: effects of zinc on steroid-binding specificity. *J. Steroid Biochem. Mol. Biol.* **85**, 195-200.

Hammond G. L. (1995). *Trends Endocrinol. Metab.* **6**, 298-304.

Hansson, T., Marelius, J., and Aqvist, J. (1998). Ligand binding affinity prediction by linear interaction energy methods. *J. Comput. Aided Mol. Des.* **12**, 27-35.

Hogeveen, K. N., Cousin, P., Pugeat, M., Dewailly, D., Soudan, B., and Hammond, G. L. (2002). Human sex hormone-binding globulin variants associated with hyperandrogenism and ovarian dysfunction. *J. Clin. Invest.* **109**, 973-981.

Jorgensen, W. L. and Tirado-Rives, J. (1988). The OPLS potential function for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J. Amer. Chem. Soc.* **110**, 1657–1666.

Joseph, D. R. (1994). Structure, function, and regulation of androgen-binding protein/sex hormone-binding globulin. *Vitam. Horm.* **49**, 197-280.

Kellenberger, E., Rodrigo, J., Muller, P., and Rognan, D. (2004). Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **57**, 225-242.

Klebe, G., Abraham, U., and Mietzner, T. (1994). Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **37**, 4130-4146.

Koshland Jr., D. (1994). The key-lock theory and the induced fit theory. *Angew. Chem. Int. Ed. Engl.* **33**, 2375-2378.

McGaughey, G. B., Sheridan, R. P., Bayly, C. I., Culberson, J. C., Kreatsoulas, C., Lindsley, S., Maiorov, V., Truchon, J. F., and Cornell, W. D. (2007). Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model* **47**,

1504-1519.

Nisker, J. A., Hammond, G. L., Davidson, B. J., Frumar, A. M., Takaki, N. K., Judd, H. L., and Siiteri, P. K. (1980). Serum sex hormone-binding globulin capacity and the percentage of free estradiol in postmenopausal women with and without endometrial carcinoma. A new biochemical basis for the association between obesity and endometrial carcinoma. *Am. J. Obstet. Gynecol.* **138**, 637-642.

Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., Debolt, S., Ferguson, D., Seibel, G., and Kollman, P. (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **91**, 1–41.

Perola, E., Walters, W. P., and Charifson, P. S. (2004). A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **56**, 235-249.

Rapuri, P. B., Gallagher, J. C., and Haynatzki, G. (2004). Endogenous levels of serum estradiol and sex hormone binding globulin determine bone mineral density, bone remodeling, the rate of bone loss, and response to treatment with estrogen in elderly women. *J. Clin. Endocrinol. Metab.* **89**, 4954-4962.

Schrödinger Inc. (2006) Glide. Version 4.0, San Diego, CA.

Scott, A. P., Pinillos, M. L., Huertas, M. (2005). The rate of uptake of sex steroids from water by *Tinca tinca* is influenced by their affinity for sex steroid binding protein plasma. *J. Fish. Biol.* **67**, 182-200.

Sotriffer, C., Krämer, O., Klebe, G. (2004). Probing flexibility and "induced-fit" phenomena in aldose reductase by comparative crystal structure analysis and molecular dynamics simulations. *Proteins* **56**, 52-66.

Van Pottelbergh, I., Goemaere, S., Zmierczak, H., and Kaufman, J. M. (2004). Perturbed sex steroid status in men with idiopathic osteoporosis and their sons. *J. Clin. Endocrinol. Metab.* **89**, 4949-4953.

Warren, G. L., Andrews, C. W., Capelli, A. M., Clarke, B., LaLonde, J., Lambert, M. H., Lindvall, M., Nevins, N., Semus, S. F., Senger, S., Tedesco, G., Wall, I. D., Woolven, J. M., Peishoff, C. E., and Head, M. S. (2006). A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **49**, 5912-5931.

Westphal J. (1971). Steroid-protein interactions. Monographs on endocrinology, Springer **4,** 567.

2 An updated steroid benchmark set and its application in the discovery of novel nanomolar ligands of sex hormone binding globulin*

2.1 List of Authors

Artem Cherkasov, Fuqiang Ban, Osvaldo Santos-Filho, Nels Thorsteinson, Magid Fallahi, and Geoffrey L. Hammond

2.2 Introduction

The blood of vertebrates contains two high-affinity steroid-binding proteins, known as sex hormone binding globulin (SHBG) and corticosteroid binding globulin (CBG), whose steroid-binding characteristics have been studied extensively (Westphal, 1971). A group of compounds known to bind to these proteins form a popular 'steroid benchmark set' utilized in many in silico modeling studies (Tuppurainen et al., 2004; Asikainen et al., 2004; Korhonen et al., 2003; Liu et al., 2002; Tuppurainen et al., 2002; Liu et al., 2001; Polanski and Walczak, 2000; Turner et al., 1999; Robinson et al., 1999; Jain et al., 1994) including popular CoMFA (Cramer et al., 1988) and CoMSIA (Klebe et al., 1994) 3D QSAR methods.

In a series of previous reports, we have investigated the SHBG system and identified various non-steroidal ligands using several innovative in silico screening methods (Cherkasov et al., 2005a; Cherkasov et al. 2006; Cherkasov et al. 2005b).

---

* A version of this chapter, co-authored by Cherkasov, A., Ban, F., Santos-Filho, O., Thorsteinson, N., Fallahi, M., and Hammond, G. L., is published in *J. Med. Chem.* (2008). 51, 2047-2056

Abbreviations: CBG, corticosteroid binding globulin; CoMFA, Comparative Molecular Field Analysis; CoMSIA, Comparative molecular similarity index analysis; GFA, Genetic Function Approximation; MMFF, Merck molecular forcefield; PLS, partial least squares; QSAR, Quantitative Structure-Activity Relationships; SHBG, sex hormone binding globulin;

Furthermore, we tested the suggested lead compounds experimentally with tritium-labeled 5α-dihydrotestosterone in a competitive ligand-binding assay. For the purpose of the current study we have combined the available data on known SHBG ligands (both steroidal and non-steroidal) and formed an expanded set of 84 compounds (shown in Table 2.1). This updated benchmark set has been used to develop various QSAR solutions enabling the discovery of non-steroidal SHBG binders.

2.3 Results

2.3.1 Docking the benchmark steroids to SHBG

We have considered the 84 SHBG ligands (including 21 steroids present in the original benchmark set) as the training set and docked all molecules into the SHBG active site using the Glide program with the default settings of the Extra Precision mode (Schrodinger Inc., 2006). For this purpose, the structure of the protein with co-crystallized ligand 5α-androstane-3β,17α–diol corresponding to the 1LHN entry of the Protein Databank was pre-optimized with the MMFF forcefield (Halgren, 1999). The ligand was then removed and the protein structure was used in the self-docking analysis, which demonstrated that the crystallographic pose of the bound steroid could be accurately reproduced (Figure 2.1).

Importantly, the Extra Precision Glide docking protocol reproduced the optimal orientation of androgens and estrogens in the SHBG steroid-binding site, in accordance with recent crystallographic and mutation experiments: specifically, C18 estrogens and C19 androgens have been shown to reside within the SHBG steroid-binding site predominantly in opposite orientations (Grishkovskaya et al., 2002a; Hammond et al., 2003). Thus, while a critical ligand-anchoring residue Ser42 in SHBG (Grishkovskaya et al., 2002a) coordinates the 17β-hydroxyl of estrogens, the same residue forms a

hydrogen bond with functional groups at the C3 position of androgens. This is illustrated in Figure 2.2, which shows the positions of 5α-dihydrotestosterone (green) and estradiol (yellow) identified in the structures of SHBG co-crystallized with these steroids (PDB entries 1KDM and 1LHU, respectively).

Notably, these crystallography-derived orientations of estrogens and androgens within the SHBG steroid-binding site (confirmed by our docking experiments) differ from the field-similarity based alignment of SHBG ligands used in the original CoMFA (Cramer et al., 1988) and CoMSIA (Klebe et al., 1994) studies, as well as in all subsequent QSAR reports involving the steroid benchmark set.

Out of 84 docked compounds 9 estrogen derivatives (*i.e.* C18-steroids containing aromatic ring "A"), i.e., estradiol, estriol, estrone, 2-iodo-estradiol, 2-metoxy-estradiol, equilenin, equilin, 17-deoxiestrone, and estradiol-3-benzoate all favored a binding pose allowing the coordination of functional groups at C17 with the Ser42 side chain. In addition, one C19 steroid, etiocholanolone was also docked in such orientation that may be attributed to some structural features of the compound (such as an unusual bending angle of a scaffold) or an artifact of the docking experiment. Otherwise, all other non-estrogen derivatives (compounds without aromatic "A" steroidal ring) demonstrated an opposite docking orientation corresponding to Ser42 anchoring functional groups at C3 of the steroid scaffold.

In addition to the correct identification of different binding modes for C18 estrogens and C19 androgens within the SHBG active site, the virtual screening protocol we have utilized produced docking scores that generally corresponded to the experimental SHBG binding constants. The resulting linear dependences with r^2=0.34 for the entire set of 84 molecules, and r^2=0.48 for the 21 benchmark steroids alone, are

shown in Figure 2.3 (it should be noted that such modest correlation coefficients are common for docking scoring functions; for instance see [Warren et al., 2006]).

Thus, we conclude that the extra precision docking of known SHBG ligands resulted in binding poses that generally reflect the preferred orientations of specific steroid classes in the crystal structures. Using the resulting docking poses of the canonical benchmark set of 21 steroids, as well as docking poses of the updated set of 84 SHBG ligands we then created CoMFA and CoMSIA models along with several additional QSAR solutions for in silico lead discovery.

2.3.2 CoMFA models

As described above, we adopted the docking poses of SHBG ligands as the basis for their structural alignment. Using this alignment (which takes into account opposite direction of SHBG binding for C18 and C19 steroids) we assembled two training sets: one corresponding to the original benchmark set of 21 steroids and an updated set including all 84 SHBG binders.

Utilization of the 1LHN-docking poses of 21 benchmark steroids resulted in a CoMFA model very similar in its statistical characteristics to that in the original study (training statistics are shown in table 2.2). This similarity in the statistical parameters of CoMFA models irrespective of drastic differences between underlying alignment shows that CoMFA lacks the ability to account for the nuances of molecular alignment. Obviously, the two models, i.e., the “classical” and the one developed here using a different (based on docking orientations) compound alignment are associated with completely different steric and electrostatic fields suggesting quite different avenues for structural modifications that should putatively lead to more active compounds. This ambiguity of CoMFA model interpretation of statistically indistinguishable alternative

models should be kept in mind as potential source of misleading hypotheses concerning novel compound design.

The CoMFA model created on the basis of the expanded set of 84 SHBG binders produced very similar training r^2 and q^2 values, but allowed 1.8-fold better recovery of the most active compounds from the training set. The corresponding Enrichment Factor (EF) values calculated with 'top 15% hit-list' criteria applied to the predicted training set values are also given in table 2.2 (more details on the EF calculations can be found in section 2.5.6).

2.3.3 CoMSIA analysis of the datasets

We also utilized the training sets of 21 and 84 superimposed molecules to create CoMSIA models and conduct comparative analysis of their accuracy and enrichment performance. Using the corresponding sets of aligned structures, we computed the standard CoMSIA fields (steric, electrostatic, hydrophobic, hydrogen bond donor, and hydrogen bond acceptor) and created PLS (Stahle and Wold, 1983) based solutions.

On one hand, the results from Table 2.2 indicate the CoMSIA solutions derived from the original benchmark and expanded sets of SHBG ligands have similar training accuracy with r^2 in 0.91-0.99 range and q^2 neighboring 0.5. On the other, as in the case of CoMFA models, the CoMSIA solution created on the bases of the expanded set of SHBG ligands allowed much better enrichment of the training set (EF=5.0), when compared to the model customized for 21 highly similar original benchmark structures (EF=2.0). Notably, the use of crystallography-complying and traditional, field similarity-based alignments of 21 benchmark steroids, as in case of CoMFA, did not result in substantially different QSAR models (both solutions are featured in Table 2.2).

2.3.4 Application of the LFER principle to protein-ligand interactions using 4D 'inductive' descriptors

Previously, 'inductive' descriptors were successfully adopted for QSAR modeling of SHBG ligands. These inductive 3D-QSAR solutions have been derived from the LFER (Linear Free Energy Relationships)-based equations for inductive and steric substituents parameters (see [Cherkasov, 2005] for more details):

$$Rs_{G\to j} = \alpha \sum_{i\subset G, i\neq j}^{n} \frac{R_i^2}{r_{i-j}^2} \qquad (1)$$

$$\sigma^*_{G\to j} = \beta \sum_{i\subset G, i\neq j}^{n} \frac{(\chi_i^0 - \chi_j^0)R_i^2}{r_{i-j}^2} \qquad (2)$$

where R_S – is the steric influence of a group of *n* atoms constituting a group *G* onto a single atom *j* (reaction centre), σ* - the inductive effect of G onto reaction center *j*. *R* corresponds to the covalent atomic radii of an *i-th* atom of a group *G*, *r* – distance between atoms *i* and *j*, χ^0 – atomic electronegativity. Parameters α and β in equations (1) and (2) normalize them to the format of Taft's original electronic and steric substituent constants (Cherkasov, 2005; Cherkasov et al., 1996).

Considering the initial success of inductive descriptors in QSAR, we adopted the LFER methodology to describe protein-ligand interactions, and updated the scope of equations (1) and (2) to the effects translated by *N*-atomic ligand *L* onto a given receptor atom *j*:

$$Rs_{L\to p} = \sum_{i\subset L}^{N} \frac{R_i^2}{r_{i-p}^2} \qquad (3)$$

$$\sigma^*_{L\to p} = \sum_{i\subset L}^{N} \frac{(\chi_i^0 - \chi_p^0)R_i^2}{r_{i-p}^2} \qquad (4)$$

where parameters $Rs_{L\to p}$ and $\sigma^{*}_{L\to p}$ describe the overall inductive and steric interactions occurring between the entire bound ligand and a receptor's atom considered as a reaction center.

Since the LFER principle is not *a priori* limited to ligand-based considerations, it is reasonable to consider how inductive and steric effects influence it in the context of inter-molecular interactions. The exact nature of inductive effects (2) is still debated, but direct electrostatic interactions (applicable for both intra- and inter-molecular levels of approximations) are often viewed as the main mechanism of its transduction (Cherkasov et al., 1996). Similarly, the model of frontal steric effects underlying equation (1) is not dependent on atomic connectivity or grouping, and can be easily applied for quantification of protein-ligand mutual screening (Cherkasov et al., 1996).

Thus, we have applied the equations (3) and (4) to the 84 compound training set placed inside the 1LHN active site, and calculated normalized R_S and σ* parameters for their optimal target binding orientations. We have considered all protein atoms (except non-polar hydrogens) in 7.0 Å ligand proximity (see section 2.5.3 for details).

In our opinion, these molecular parameters $Rs_{L\to p}$ and $\sigma^{*}_{L\to p}$ can be regarded as receptor-dependent 3D-QSAR descriptors because they are derived from three-dimensional structures of compounds, and rely on their positioning within the target protein taking into account pairing of protein and ligand atoms. We expected that such 4D inductive descriptors would possess good predictive ability and illustrate the advantage of using correct steroidal alignment in modeling SHBG binding.

2.3.5 GFA-based QSAR models

Considering all structures from the training set (84 entries), we computed the full spectrum of R_S and σ* values (one of each for every considered atom of 7.0 Å ligand

surrounding). To relate such a large number of descriptors to dependent variables pK_d we employed the Genetic Algorithm approximation, which has been applied for QSAR analyses relying on a heuristic search (Holland, 1975).

In our current study, we adopted the Genetic Function Approximation (GFA) developed by Rogers and Hopfinger, 1994, which is based on the G/SPLINES Genetic Algorithm implementation (Rogers, 1991; Rogers, 1992). Given a large number of QSAR descriptors to sample, this approach creates a 'population' of QSAR models and applies the 'fitness function' to iteratively evolve them to an optimal solution (i.*e.* to find the most appropriate set of descriptors). The GFA approach uses the Friedman's 'lack-of-fit' (LOF) fitness criteria:

$$LOF = \frac{LSE}{\left(1 - \frac{c + dp}{n}\right)^2}$$

where *LSE* is the least squared error; *c* is the number of descriptors employed by the model; *d* is the user-defined smoothing factor, *p* is the total number of available descriptors; and *n* is the number of the training set molecules.

We have applied the GFA approach implemented within the WOLF package (Chem21 Group, Inc., 2007) with the following default settings: the initial population of QSAR models has been limited to 5 000; the total number of crossovers was set to 200 000 and up to 50 % of models were allowed to mutate in every generation (i.e. to randomly sample descriptor values). The resulting linear QSAR solutions have been constructed using the PLS approach and have been further validated by the leave-one-out (LOO) procedure.

The parameters of the final optimal QSAR solution based on six inductive descriptors are presented in Table 2.2, while the predicted activity parameters are listed

in Table 2.1. As these results indicate, the Genetic Algorithm provided modestly accurate but, nonetheless, reasonable and simple models predicting 84 SHBG binding constants with correlation coefficients r^2= 0.56 and q^2 = 0.45. The developed GFA-QSAR model could efficiently rank the most active ligands and despite modest training statistics allowed a 3.5-fold hit enrichment.

2.3.6 'QuaSAR-Evolution' models

In addition to the GFA method, we used the Genetic Algorithm-based approach implemented by the *QuaSAR-Evolution* module of the MOE program (Chemical Computing Group, Inc., 2006). This tool enables automated QSAR modeling 'on the fly' and is available through the 'SVL exchange' (Chemical Computing Group, Inc., 2005).

We applied the *QuaSAR-Evolution* tool with its default settings*:* (a) the initial population of 100 models; (b) 4 additional descriptors added to each generation of QSAR models; (c) multiple linear regression (MLR) mode; (d) the total number of crossovers set to 50 000; (e) allowed 50 % mutation and (f) the auto-termination factor of 1 000 (meaning that the calculation was stopped when the 'fitness function' value does not change during 1 000 crossovers). The resulting QSAR solution demonstrated better training accuracy compared to the previous GFA models with r^2=and q^2 estimated as 0.66 and 0.58, respectively (also allowing 3.50-folds enrichment of the training set). It should also be mentioned that the developed linear QSAR model could provide some insight into factors determining SHBG dissociation constants:

$$pK_d = 4.63 - 0.98Rs\_ASP65 - 1.18Sigma\_ASP65 + 0.36Sigma\_GLY58 + 1.22\ Sigma\_HIS136 - 0.41Sigma\_LEU69 + 0.66Sigma\_PHE44 - 0.54\ Sigma\_THR60 - 0.65Sigma\_VAL127; \quad (5)$$

N = 84 r^2 = 0.66 q^2 = 0.58 SE = 0.79

The above LFER equation illustrates that one of the most significant contributions to pK_d comes from inductive interaction between a ligand and the Asp65 side-chain (critical anchoring residue located at the 'gating' mobile loop of the SHBG active site believed to control ligand uptake and release [Hammond et al., 2003]). Steric interactions with Asp65 that are described by the Rs_ASP65 descriptor also play an important role in ligand binding, such that (5) illustrates its minimization helps to increase pK_d.

Similarly, according to (5) the electron-withdrawing effect exhibited toward Gly58 should also increase pK_d, and it is likely that the backbone oxygen of that residue influences the polar stabilization and hydrogen bonding of some SHBG ligands. The possible role of inductive interactions between a bound ligand and Thr60 is also reflected by (5).

Of note, the involvement of residues His136, Leu69, Phe44 and Val129 featured in (5) with respect to protein-ligand interactions is less obvious, and some known ligand-binders such as Ser42 are not reflected by (5). These inconsistencies can perhaps be explained by the limited variability of inductive and steric effects of some residues or by the approximate nature of Genetic Algorithm solutions.

Overall, the above results substantiate the adequate accuracy of the developed QSAR models and their ability to account for specific protein-ligand interactions.

2.3.7 Consensus scoring by the developed QSAR models

To further expand the utility of the developed QSAR models we have implemented the consensus scoring approach. Thus, eight different predicted parameters of potential activity (two Glide, two CoMFA, two CoMSIA, one *GFA-QSAR* and one *QuaSAR-Evolution* values) have been produced for every entry in the training

set. Based on these sorted values, each molecule would then receive a binary 1.0 vote for every 'top15 % appearance' (thus, the maximum possible vote was set to 8.0). The final cumulative vote was then used to rank the training set entries.

The resulting 5.0 fold enrichment of the top 15 % binders in the "hit list" clearly demonstrated that the consensus scoring strategy produces the most balanced predictions, and that a synergistic approach can capitalize on the strengths of individual approaches (such as the positive predictive power of ligand-based QSAR techniques and the negative predictive power of docking) and compensate for their weaknesses. Furthermore, we have evaluated several other combined strategies (also featured in Table 2.2) and discovered that all of them resulted in consistent 4.0-5.0 folds enrichments of the training set.

It should be noted that the use of complementary predicting tools and the implementation of scoring/voting protocols has recently become one of the most important topics in the field of computer-aided drug design (Feher, 2008).

2.3.8 Selection of potential SHBG binders

Overall, the results of QSAR modeling of the training set resulted in good accuracy and their synergy resulted in the ability to enrich for the most active target binders. These observations encouraged us to apply our scoring systems to electronic collections of commercially available chemicals for the identification of novel non-steroidal SHBG binders. In this study we used the ZINC 5.0 molecular database (Irwin and Shoichet, 2005) that included 3.3 million entries. From these, we derived 2 066 886 non-redundant molecules satisfying drug-likeness criteria (see section 2.5.1 for details).

As described in the previous section, all created fields-based and 4D-QSAR solutions were based on high-precision docking poses. Therefore, in order to apply pre-

trained CoMFA, CoMSIA and 4D-QSAR models we docked all 2 066 886 structures into the 1LHN ligand-binding site. This protein structure was used because, as previously noted, it allowed us to produce the correct docking poses and binding characteristics of the training set compounds. Furthermore, to account for possible induced changes in the SHBG active site, we also docked all 2 066 886 molecules into the 1KDM protein structure which contains SHBG co-crystallized with 5α-dihydrotestosterone, the compound with the highest binding affinity (Grishkovskaya et al., 2002b). All molecular structures that produced GlideScore values <-7.0 were selected and thus, two redundant hit-lists have been generated, with one of them corresponding to the 143 421 best 1KDM-docked ligands and the other corresponding to the top 213 191 1LHN-predicted binders. Next, we implemented a scoring system that assigned a 1.0 vote to the top 5 % of both 1KDM (7171) and 1LHN (10659) hits, while all other docked ligands were given a vote value of 0. Based on the resulting cumulative vote we selected 3759 structures for future assessment.

All of the selected docking poses were examined visually; several broken and inconsistently docked structures were removed, and all steroidal derivatives and compounds containing a carboxylic group were eliminated, in order to reduce the total number of selected structures to 1419. All of these ligands were then re-docked into the 1KDM and 1LHN active sites using the XP_Glide (Schrödinger, Inc., 2006).

The resulting 'extra precision' docking poses were scored by CoMFA_21, CoMFA_84, CoMSIA_21, CoMSIA_84 and *QSAR_GFA, QuaSAR-Evolution* solutions, where '_21' and '_84' symbols mark models created on the original and updated sets of SHBG ligands respectively.

After sorting all eight sets of predicted activities, we computed the cumulative votes for 1419 molecular structures, where an entry would receive a vote for every 'top 15 %' appearance. Based on these cumulative parameters (with the maximal possible value of 8.0), we selected a list of 111 hits, all of which had been voted on 4 or more times.

After the final visual inspection, we formed a list of 87 compounds out of which 41 chemical substances could be readily purchased in sufficient purity and quantity for biochemical verification as SHBG binders, as described below.

2.3.9 Experimental testing

All 41 compounds selected from the 3.3 million ZINC entries by applying drug-likeness criteria followed by the combination of docking, CoMFA, CoMSIA and 4D QSAR filters, were further screened for their ability to interact with the SHBG steroid-binding site *in vitro*. The screening assay involved a modification of an established competitive steroid ligand-binding assay that employs tritium labeled 5α-dihydrotestosterone as the radio-labeled ligand (see section 2.5.7 for details).

The initial experimental screen of test compounds was conducted at a single high concentration (approximately 100 μM), and 25 out of 41 compounds demonstrated some SHBG-binding competition with the tritium labeled 5α-dihydrotestosterone. The seven substances that displaced more than 50 % of bound 5α-dihydrotestosterone from the protein at the 100 μM concentration were further analyzed in a concentration-dependent manner (pK_d values and data on all compounds tested for SHBG binding can be found in Table 2.1).

The competitive displacement curves generated using these non-steroidal SHBG ligands (entries **85-91** in Table 2.1) are presented in Figure 2.4, and the corresponding SHBG dissociation constants calculated from the plot are included in Table 2.1.

The 5 most active SHBG ligands exhibited nanomolar dissociation constants: 106.9 nM for compound **85**, 408.8 nM for compound **86**, 591.3 nM for compound **88**, 833.6 nM for compound **89** and 964 nM for compound **90**. The parallelism of the corresponding competitive displacement curves in Figure 2.4 indicates that these compounds are completely soluble at high concentrations and behave in essentially the same way as a steroid ligand with respect to their kinetics of binding. Taken together with their high affinity toward the target, these novel SHBG binders represent potential therapeutic prototypes.

It is also worth mentioning, that this hit rate appears very good (25 out of 41 tested chemicals showed some activity, with 7 of them being nanomolar to low micromolar binders), especially considering that we could only test available (as opposed to custom-made) compounds, and taking into account the financial constraints of academia-based drug discovery research.

2.3.10 Non-steroidal SHBG binders

Analysis of the docking poses of the 8 most active ligands (for 7 of them pK_d values could be measured) provided additional and important insight into the mechanism of SHBG binding. As Figure 2.5 illustrates, all 8 ligands likely form a hydrogen bond with the Asp65 side chain (supported by a secondary interaction with Asn82), with two of them, compounds **89** and **90,** also showing strong hydrogen bonding toward Ser42 (supported by additional interaction with Val105 backbone oxygen). These observations illustrate the importance of the Asp65 and Ser42

anchoring residues previously outlined in numerous SHBG-related publications (Grishkovskaya et al., 2002a; Hammond et al., 2003). It should be noted, however, that the presence of two anchoring hydrogen bonds did not make the compounds better binders, in fact, three other substances (**85**, **86** and **88**) all formed only one hydrogen bond, but demonstrate higher affinity toward the target (likely caused by more favorable hydrophobic interactions).

The importance of hydrophobic forces is well recognized for SHBG binding (Grishkovskaya et al., 2002a) and it is therefore no surprise that all 8 ligands have sizable aliphatic and aromatic cores that could participate in close-range interactions within hydrophobic pockets. One such pocket is located in close proximity to the Ser42 residue and is formed by the Leu171, Met138 and Val105 side chains of hSHBG. The latter residue also forms a hydrophobic patch together with the Phe67 side chain providing additional stabilization for bound ligands. Importantly, Phe67 is also likely involved in π-stacking with aromatic rings of **88**, **89** and **90** and perhaps with the C=O group of **87** (see Figure 2.5 for more details). It is also possible that the SHBG affinity for compounds **85**, **86** and **88**, which involves strong hydrophobic interactions with key residues within its steroid-binding pocket, could be further increased by introducing additional H-bond enabling groups into their Ser42-oriented ends. Such structural modifications could represent an attractive strategy for lead optimization. It is also possible that an extra hydrogen-bond acceptor to the Asp65-oriented end of a ligand that would engage the Asn82 side chain could enhance binding, as would appear to occur in the cases of compounds **88** and **90**.

Another ‘atypical’ interaction within the active site has been found for ligand **91,** which forms an additional H-bond with the Gly58 backbone oxygen. Such coordination

has never been previously observed for SHBG ligands, but the possible relevance of this residue was hinted at by the LFER-equation (5).

2.4 Conclusions

Using available information on known ligands of SHBG we have developed several structure-activity models based on conventional (CoMFA and CoMSIA) and newly developed QSAR approaches. While building such QSAR solutions we used molecular alignments that contradict the conventional way of superimposing steroidal SHBG ligands, but are in line with direct crystallographic evidence of the steroid-binding poses. We have demonstrated that molecular-field based techniques such as CoMFA and CoMSIA are not very sensitive to ligand alignment, as they result in almost indistinguishable QSAR models derived from the traditional and 'crystallographic' alignments of steroidal scaffolds.

We developed novel ligand-induced active site descriptors (called inductive 4D QSAR parameters) which provided additional insights into factors influencing ligand-protein interactions and which have been successfully used in combination with other in silico drug discovery tools. Thus, the developed range of in silico solutions have been applied in a consensus manner to more than 2 million structures from the ZINC database and identified 41 potential SHBG binders, 25 of which demonstrated detectible binding to SHBG in plasma. Notably, 5 such novel non-steroidal SHBG inhibitors demonstrated nanomolar dissociation constants, with the best binder exhibiting K_d=109 nM and representing the most active non-steroidal SHBG ligand known to date.

Since SHBG represents a prospective drug target the identified non-steroidal lead compounds can be characterized as potential therapeutic agents laying a foundation for future lead optimization studies.

2.5 Materials and Methods

2.5.1 Database preparation

The initially considered set of 3.3 millions compounds from the ZINC 5.0 database was reduced to 2 066 886 entries by applying the drug-likeness criteria: molecular weight between 300 and 800 Da; the presence of 1 to 10 hydrogen bond acceptors; 1 to 5 hydrogen bond donors; less than 10 rotatable bonds, and overall hydrophobicity below log P = 5.00.

The resulting set of 2 066 886 drug-like structures has been washed – *i.e.* all inorganic components have been removed, and all ionizable groups have been coordinated with pH=7.0 conditions.

All molecular structures have been optimized using PM3 semi-empirical method implemented within MOE.

2.5.2 Docking

The Maestro suite (Schrodinger Inc., 2004) was used to prepare the 1LHN and 1KDM protein structures for docking. All water and ion molecules were removed from the corresponding PDB files, and hydrogen atoms were added and adjusted where necessary. Steroid-binding sites were defined as 10 Å surrounding the co-crystallized ligands in 1KDM and 1LHN.

2.5.3 QSAR descriptors calculation and model building

The extra precision docking poses of 84 training set compounds, 64 chemicals investigated in our previous SHBG studies and 1419 selected molecules were used to

compute 4D inductive QSAR descriptors according to (3) and (4). For every docked molecule placed in the 1LHN active site (defined as 7 Å surrounding of its native ligand) we computed the direct 3D distances from all atoms of a ligand to the active site's polar hydrogens and heavy atoms. The computed distances were used in (3) and (4) to compute the cumulative parameters of inductive σ* and steric R_S influence of a ligand to every considered atom within the active site. All calculations were implemented with customized SVL scripts of MOE.

The defined 1LHN active site contained 289 heavy atoms and polar hydrogens and, therefore, for every ligand we computed 2* 289 = 578 descriptors. These parameters were then used to create predictive QSAR solutions based on the Genetic Algorithm approximation.

The *QuaSAR-Evolution* models were built using the 'autoqsar.svl' script obtained from the SVL exchange site. The default setting was used.

The GFA models were created using WOLF 6.2 software (with default settings) kindly provided by Professor Hopfinger (Chem21 Group Inc., 2007). Both of these programs automatically handle the descriptors' cross-correlation problems and possess built-in capabilities for LOO cross-validation.

The actual values of normalized 4D QSAR descriptors can be obtained upon request.

2.5.4 Molecular alignment

For generating the 'traditional' set of superimposed steroidal structures we used the SYBYL (Tripos Inc., 2006) *Fit_Atoms* functionality, which is based on the BMFIT method (Nyburg, 1974). The 1LHN docking pose of its ligand was used as the reference, while other molecules were translated and rotated in a way to fit the weighted

centroid of atoms of the “A” ring of steroidal scaffolds. In this way, all steroids have been superimposed in ‘head-to-head’ orientations, and all non-steroidal structures were also taken in their docked configurations.

In the ‘corrected’ datasets all steroidal and non-steroidal ligands were used in their respective 1LHN docking poses.

2.5.5 CoMFA modeling

SYBYL was used to construct all CoMFA models using the partial least squares fitting, with the cross-validation carried out by the built-in LOO procedure.

Both the traditional and expanded data sets of SHBG binders (containing 21 and 84 entries, respectively) were used independently to compute steric and electrostatic CoMFA fields. The steric ones were calculated on 2 Å grids, by evaluating ‘6-12’ Van der Walls interaction with default CoMFA probes. We used distance-dependent dielectric parameters to compute the Coulombic interactions approximating electrostatic CoMFA fields, and set the field’s truncation parameter to 30.0 kcal/mol.

For the traditional 21 steroids of the benchmark set we also re-created CoMFA models based on traditional similarity-based molecular alignment used in, with the resulting statistics reproducing original values reported in (Cramer et al., 1988).

2.5.6 CoMSIA modeling

For the studied datasets, we computed 5 CoMSIA properties that included steric, electrostatic, hydrophobic, hydrogen bond donor, and hydrogen bond acceptor fields (computed with default settings). The fields were derived according to similarity indexes (computed with 0.3 attenuation factor) of molecules brought into a common alignment. In the CoMSIA study, we utilized the same alignment schemes as in CoMFA modeling.

All calculations were carried out with default settings; each CoMSIA property of a given atom was scaled to 74.1 % for its 1 Å proximity, to 30.1 % for >2 Å surrounding, and to 6.7 % for the area within 3 Å.

The final CoMSIA models were constructed using the partial least squares (PLS) algorithm (Stahle and Wold, 1988) and cross-validated by the LOO procedure implemented by the SYBYL package.

2.5.7 Enrichment calculations

All predicted SHBG affinity parameters (including Glide docking scores, CoMFA, CoMSIA and 4D-QSAR outputs) have been processed into the parameters of percent yield (%Y), percent accurate (%A), enrichment factor (EF) and goodness of hit list (GH) parameters custom for in silico screening studies:

$\%Y = H_a/H_t$

$\%A = H_a/A$

$EF = (H_a/H_t) / (A/D)$

$$GH = \left(\frac{H_a(3A + H_t}{4H_t A} \right) \times \left(1 - \frac{H_t - H_a}{D - A} \right)$$

Where H_t is the total number of compounds in the hit list (in our case – top 15 % portion of the sorted predictions), H_a is the number of known actives in the hit list (true positives), A is the active compounds in the database, D is the number of compounds in the database. We have only reported the EF values; other parameters can be obtained from authors upon request.

The corresponding calculations have been carried out using in-house SVL scripts.

2.5.8 SHBG ligand-binding assay

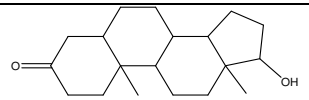
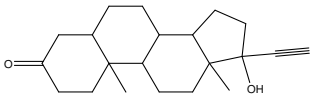
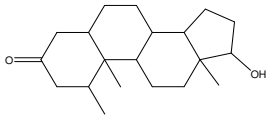
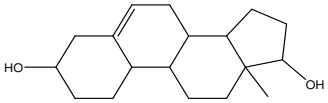
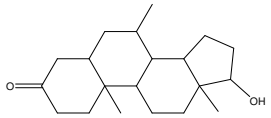
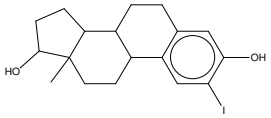
An established competitive ligand-binding assay was used to determine the relative binding affinities of the studied compounds to human SHBG, compared to testosterone and estradiol standards (Hammond and Lahteenmaki, 1983). In brief, the assay involved mixing 100 µl aliquots of diluted (1:200) human pregnancy serum containing approximately 1 nM SHBG, which was pre-treated with dextran-coated charcoal (DCC) to remove endogenous steroid ligand, with 100 µl of tritium labeled 5α-dihydrotestosterone at 10 nM as labeled ligand. For the screening assay, triplicate aliquots (100 µl) of a fixed amount (100 µM) of test compound was added to this mixture and incubated overnight at room temperature. After further 10 min incubation at 0 ºC, 500 µl of a DCC slurry was added at 0 ºC, and incubated for 10 min prior to centrifugation to separate SHBG-bound from free 5α-dihydrotestosterone. Compounds that displaced more than 35 % of the tritium labeled 5α-dihydrotestosterone from the SHBG in this assay were then diluted serially, and triplicate aliquots (100 µl) of known concentrations of test compounds were used to generate complete competition curves by incubation with the SHBG/5α-dihydrotestosterone mixture, and separation of SHBG-bound from free steroid, as in the screening assay. The amounts of 5α-dihydrotestosterone bound to SHBG at each concentration of competitor ligand were determined by scintillation spectrophotometry and plotted in relation to the amount of 5α-dihydrotestosterone bound to SHBG at zero concentration of competitor. From the resulting competition curves, IC_{50} concentrations could be calculated if displacement of more than 50 % of tritium labeled 5α-dihydrotestosterone from SHBG was achieved.

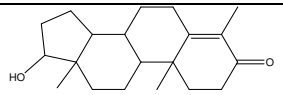
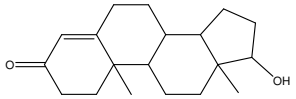
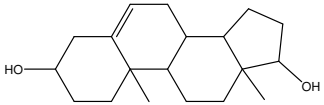
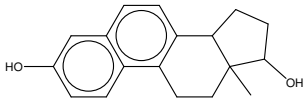
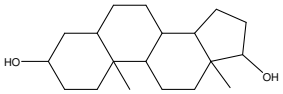
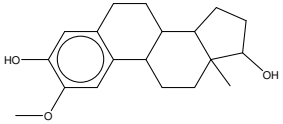
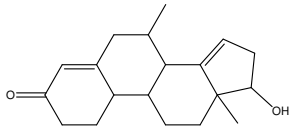
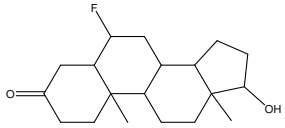
The dissociation constants (K_d) have been calculated from the relative binding affinity parameters using the following equation: $1/\{K_a(\text{dihydrotestosterone})/[(1+R)/RBA$

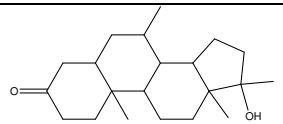
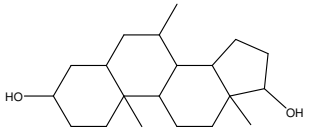
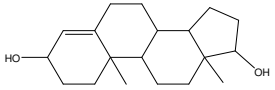
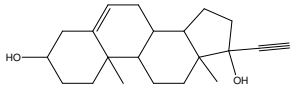
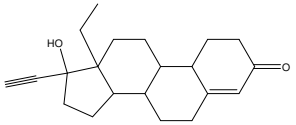
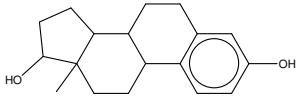
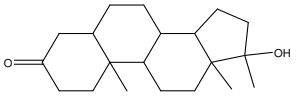
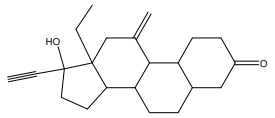
- R]}, where K_a(dihydrotestosterone) = 0.98 * 10^9 M^{-1} is the association constant of the 5α-dihydrotestosterone and R (0.05) is the ratio of bound-to-free 5α-dihydrotestosterone at 50 % displacement in the assay.

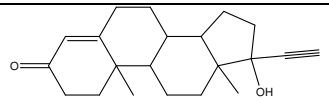
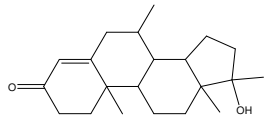
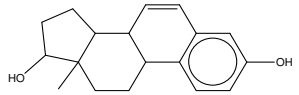
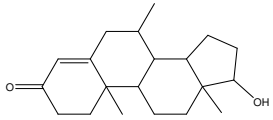
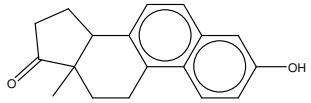
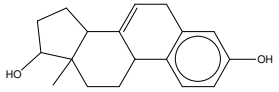
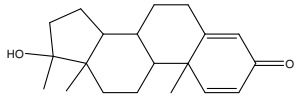
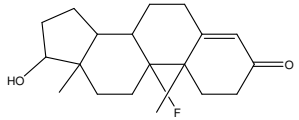
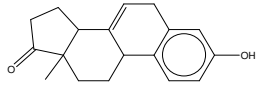
Table 2.1

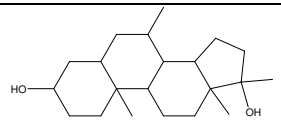
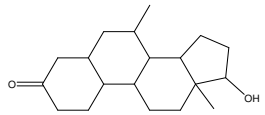
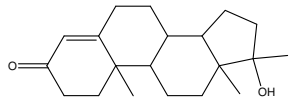
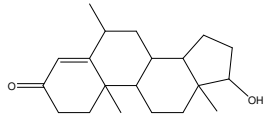
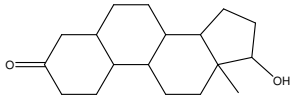
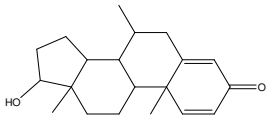
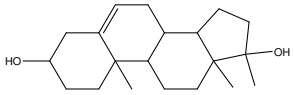
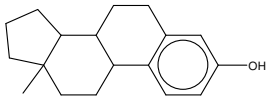
Known SHBG binders utilized for QSAR modeling and novel non-steroidal ligands identified in the current study

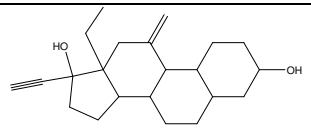
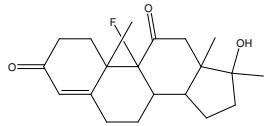
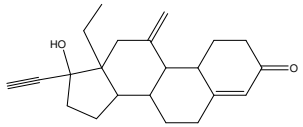
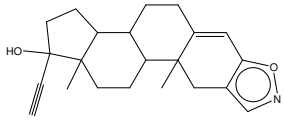
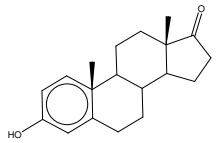
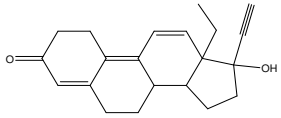
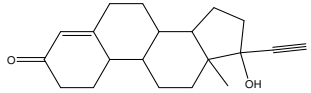
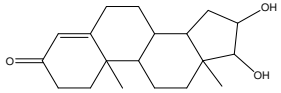
#	Name	Structure	pK_d	I	II	III	IV	V	VI	VII	VIII
TRAINING SET											
1	5α-Dihydrotestosterone [#]		**9.74**	8.97	8.67	9.71	9.29	9.57	8.79	-13.23	-14.8
2	17-ethinyl-Dihydrotestosterone		**9.74**	9.24	9.37	7.94	9.38	9.03	8.70	-12.8	-12.52
3	1α-methyl-Dihydrotestosterone		**9.60**	8.87	8.76	8.95	9.62	9.09	8.99	-13.75	-15.04
4	5-Androstene-19-nor-3b,17β-diol		**9.54**	8.68	8.84	8.67	8.69	8.88	8.66	-13.38	-13.78
5	7α-methyl-Dihydrotestosterone		**9.38**	8.97	8.76	8.94	9.65	9.07	9.00	-13.69	-15.11
6	2-Iodo Estradiol		**9.32**	8.68	7.46	6.58	9.62	7.24	9.33	-10.56	-11.14

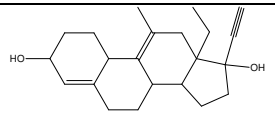
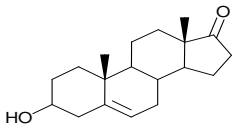
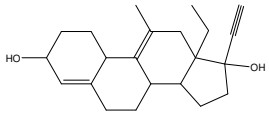
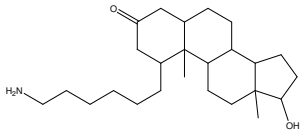
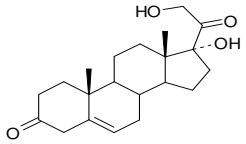
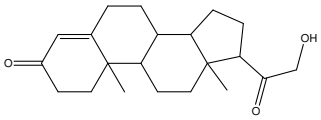
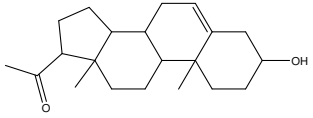
#	Name	Structure	pK_d	I	II	III	IV	V	VI	VII	VIII
7	4-methyl-Testosterone		**9.31**	9.23	8.94	8.86	9.36	9.14	9.54	-14.14	-15.23
8	Testosterone[#]		**9.20**	8.60	8.65	9.23	9.23	9.23	8.97	-13.89	-14.94
9	5α-Androstene-3β,17β-diol[#]		**9.17**	8.76	8.47	9.24	9.34	9.30	9.04	-13.28	-13.4
10	Dihydroequilenin-17β		**9.12**	8.65	8.14	10.40	9.05	9.05	8.42	-13.72	-13.36
11	5α-Androstane-3α,17β-diol[#]		**9.11**	8.85	8.76	9.10	9.64	9.24	9.25	-14.07	-13.36
12	2-Methoxy Estradiol		**9.08**	8.49	7.70	9.73	9.26	8.98	9.02	-12.45	-12.06
13	7α-methyl-14-dehydro-19-Nortestosterone		**9.07**	8.22	8.26	8.59	8.40	8.90	8.99	-12.98	-14.09
14	6α-fluoro-Dihydrotestosterone		**9.05**	8.84	8.33	8.43	8.82	9.19	8.98	-12.23	-13.92

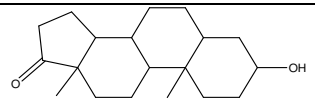
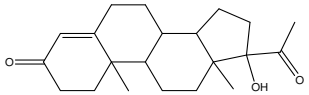
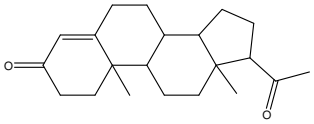
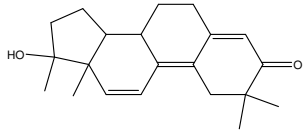
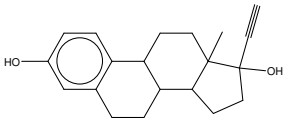
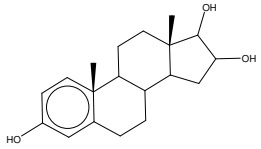
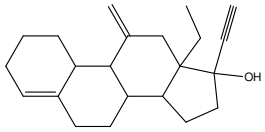
#	Name	Structure	pK_d	I	II	III	IV	V	VI	VII	VIII
15	7α-17-dimethyl-Dihydrotestosterone		**9.05**	7.89	7.34	8.79	9.10	8.86	9.40	-14.25	-14.86
16	7α-methyl-5α-Androstane-3β,17β-diol		**9.00**	7.52	7.45	8.48	8.77	9.13	8.95	-12.69	-13.89
17	4-Androstene-3β,17β-diol		**9.00**	8.85	8.86	9.39	9.02	9.01	9.05	-13.84	-15.19
18	17-ethinyl-delta, 5-Androstane		**8.91**	8.18	8.57	8.63	8.53	8.73	8.68	-13.97	-14.24
19	d-Norgestrel		**8.91**	8.44	7.35	7.76	7.89	8.75	7.92	-12.72	-13.39
20	Estradiol[#]		**8.83**	7.14	7.48	8.85	8.35	8.83	7.70	-13.74	-14.29
21	17-methyl-Dihydrotestosterone		**8.81**	7.58	8.21	8.23	9.03	9.01	9.13	-13.56	-14.67
22	17-ethinyl-11-methylene-18-methyl-19-nor-Dihydrotestosterone		**8.80**	8.66	8.33	8.40	8.69	9.23	8.82	-12.61	-12.96

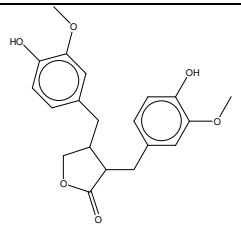
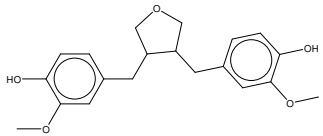
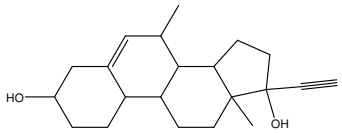
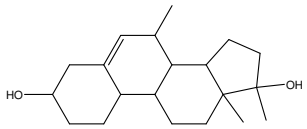
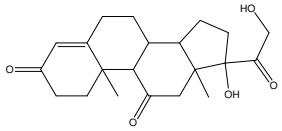
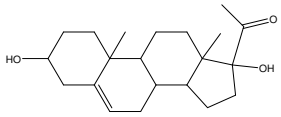
#	Name	Structure	pK_d	I	II	III	IV	V	VI	VII	VIII
23	Ethisterone		**8.78**	7.83	8.12	8.13	8.60	8.79	8.42	-13.99	-14.01
24	7α,17-dimethyl-Testosterone		**8.76**	8.37	8.57	8.00	8.85	8.96	8.88	-13.81	-12.43
25	6-dehydro-Estradiol		**8.76**	8.45	8.43	10.03	9.34	9.16	8.47	-13.51	-13.9
26	7a-methyl-Testosterone		**8.71**	8.75	8.74	8.79	8.40	9.05	9.06	-13.47	-14.34
27	Equilenin		**8.62**	8.48	8.43	10.27	8.27	9.06	8.92	-12.46	-10.99
28	7-dehydro-Estradiol		**8.62**	8.68	9.06	7.45	8.19	7.56	8.52	-13.51	-13.9
29	17-methyl-1-Dihydrotestosterone		**8.57**	7.62	7.95	8.03	8.84	9.16	8.80	-14.03	-12.95
30	9a-fluoro-Testosterone		**8.51**	9.59	8.11	8.99	8.77	8.65	8.36	-13.71	-14.22
31	Equilin		**8.51**	8.75	9.15	7.57	8.51	7.75	8.15	-11.57	-11.2

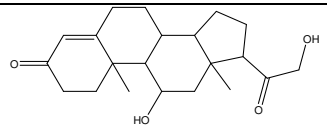
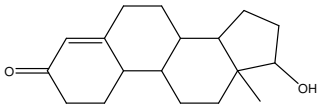
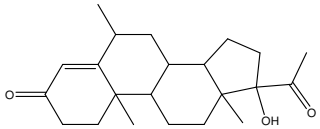
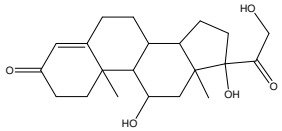
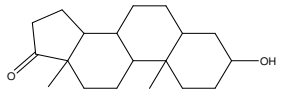
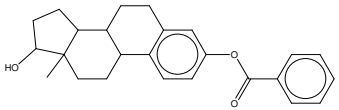
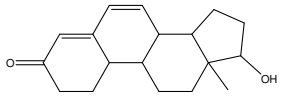
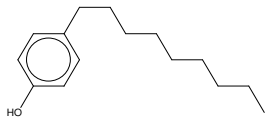
#	Name	Structure	pK_d	I	II	III	IV	V	VI	VII	VIII
32	7α,17-dimethyl-5a-Androstane-3β,17β-diol		**8.46**	8.28	9.06	7.47	8.04	9.18	8.56	-13.22	-12.03
33	7α-methyl-19-nor-Dihydrotestosterone		**8.46**	8.23	8.04	9.26	7.86	9.54	7.91	-12.75	-12.01
34	17-methyl-Testosterone		**8.43**	8.32	8.47	7.97	8.42	8.92	8.79	-13.85	-12.29
35	17-ethinyl-3,3-difluoro-5a-Androstan-17β-ol		**8.36**	7.43	6.69	8.49	8.53	9.16	9.10	-12.18	-11.9
36	6α-methyl-Testosterone		**8.36**	7.24	8.31	8.11	8.47	9.09	8.40	-11.74	-12.27
37	19-Nor-Dihydrotestosterone		**8.36**	8.12	8.66	9.08	7.70	9.42	8.31	-11.8	-13.11
38	7α-methyl-1-dehydro Testosterone		**8.36**	8.81	8.36	8.80	7.61	9.43	7.83	-13.32	-13.03
39	17-methyl-delta-5-Androstane		**8.36**	9.06	8.31	8.63	8.57	9.18	8.39	-13.68	-11.91
40	17-Deoxoestrone		**8.30**	7.42	7.77	8.61	8.67	8.65	8.33	-12.4	-13.65

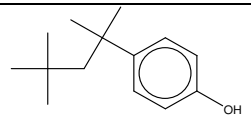
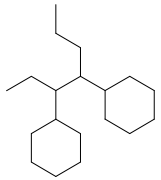
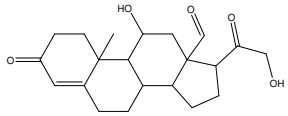
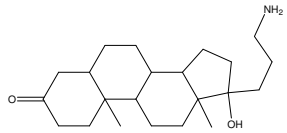
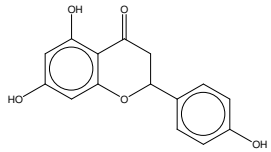
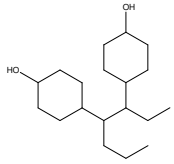
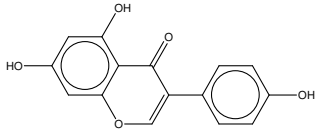
#	Name	Structure	pK_d	I	II	III	IV	V	VI	VII	VIII
41	3α-hydroxy-5α-H-17-ethinil-11-methylene-18-methyl-19-nor-Dihydrotestosterone		**8.27**	8.07	7.93	7.86	8.38	8.39	8.40	-12.38	-13.09
42	9α-fluoro-11-oxo-17-methyl-Testosterone		**8.23**	7.88	7.83	7.23	8.37	8.04	7.60	-13.1	-12
43	17-ethinyl-11-methylene-18-methyl-19-nortestosterone		**8.23**	8.34	8.41	7.60	8.71	8.71	8.25	-12.31	-13.59
44	Danazol		**8.20**	8.51	8.97	7.68	8.14	8.57	8.12	-12.85	-8.94
45	Estrone[#]		**8.18**	7.71	8.38	8.16	8.26	8.14	7.46	-12.8	-11.43
46	17-ethinyl-18-methylestra-4,9,13-trien-17β-ol-3-one		**8.11**	7.85	8.39	8.54	7.87	8.88	7.91	-13.34	-13.62
47	Norethindrone		**7.97**	7.91	7.93	8.50	7.92	8.86	7.39	-13.02	-13.83
48	16α-hydroxy Testosterone		**7.92**	7.83	7.59	7.49	7.80	6.03	8.14	-12.44	-13.83

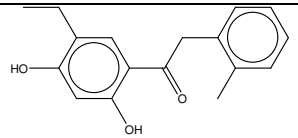
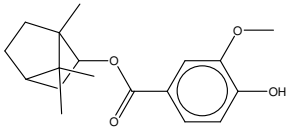
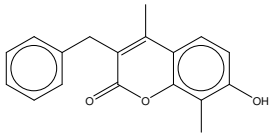
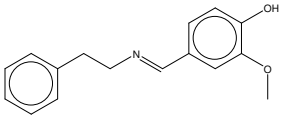
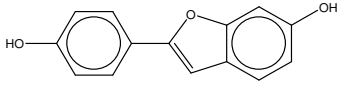
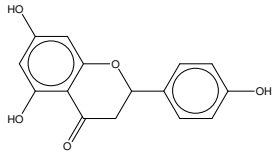
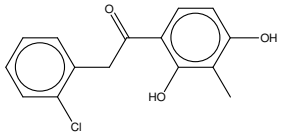
#	Name	Structure	pK_d	I	II	III	IV	V	VI	VII	VIII
49	3b-hydroxy-17-ethinyl-11-methylene-18-methyl-4-Estren-17β-ol		**7.88**	7.73	8.66	6.87	7.66	7.83	7.43	-8.86	-11.42
50	Dehydroepiandrosterone[#]		**7.84**	8.52	7.53	7.77	7.94	7.69	8.05	-12.53	-1.33
51	3α-hydroxy-17-ethinyl-11-methylene-18-methyl-4-Estren-17β-ol		**7.67**	7.86	7.67	8.90	7.25	9.45	8.26	-11.26	-12.87
52	1α-aminohexyl-17β-hydroxy-5α-Androstan-3-one		**7.53**	8.23	7.66	7.46	7.56	7.44	7.97	-11.53	-12.05
53	Androstenedione[#]		**7.46**	8.01	7.63	8.17	6.67	8.70	6.81	-11.95	-10.7
54	Deoxycortisol[#]		**7.44**[$]	7.51	7.72	7.22	6.77	7.14	7.40	-12.86	-13.29
55	Deoxycorticosterone[#]		**7.38**	6.47	6.92	7.40	7.39	7.29	7.65	-12.19	-12.14
56	Pregnenolone[#]		**7.15**	7.90	7.53	7.03	6.79	6.91	6.80	-11.38	-11.5

#	Name	Structure	pK_d	I	II	III	IV	V	VI	VII	VIII
57	Androsterone[#]		**7.15**	7.33	6.39	7.18	6.96	7.20	7.64	-13.98	-12.45
58	17-hydroxy-Progesterone[#]		**7.00**	7.37	6.77	6.86	6.65	6.87	6.98	-11.66	-13.29
59	Progesterone[#]		**6.94**	8.59	7.60	6.99	7.09	7.20	7.49	-10.72	-12.59
60	17β-hydroxy-2,2,17-trimethylestra-4,9,11-trien-4-one		**6.86**	5.94	5.86	7.84	6.15	7.64	6.22	-11.13	-9.57
61	17-ethinylestradiol		**6.81**	7.58	7.34	7.87	7.35	8.70	7.07	-13.15	-13.24
62	Estriol[#]		**6.63**	7.45	8.02	6.62	7.16	6.61	7.43	-11.92	-12.57
63	17-ethinyl-11-methylene-18-methyl-4-Estren-17β-ol		**6.60**	6.47	8.00	6.41	6.65	7.16	6.65	-12.78	-13.22

#	Name	Structure	pK_d	I	II	III	IV	V	VI	VII	VIII
64	(-)-Matairesinol		**6.51**	7.95	8.66	7.52	6.67	8.53	7.15	-10.37	-9.88
65	3,4-divanillyl Tetrahydrofuran		**6.51**	7.82	7.15	7.49	6.72	8.45	6.78	-10.99	-10.36
66	7α-methyl-17-ethinyl-delta5E		**6.44**	6.54	6.63	7.53	6.35	6.76	5.93	-12.15	-13.16
67	7α,17-dimethyl-delta5E		**6.44**	6.17	6.37	6.55	6.41	6.57	6.23	-12.44	-12.08
68	Cortisone[#]		**6.43**	8.10	6.67	6.43	6.37	6.45	6.20	-10.03	-10.07
69	17-hydroxy-5-Pregnen-3b-ol-20-one[#]		**6.36**	6.27	7.92	9.90	7.91	9.37	8.06	-11.63	-9.71

#	Name	Structure	pK_d	I	II	III	IV	V	VI	VII	VIII
70	Corticosterone[#]		**6.34**	7.47	6.55	6.35	6.73	6.42	7.05	-10.37	-9.05
71	19-Nor-Testosterone		**6.30**	6.04	6.11	6.18	6.34	6.22	5.98	-12.21	-13.28
72	17-hydroxy-6α-methyl Progesterone		**6.20**	8.42	7.97	6.14	5.85	6.10	6.19	-9.68	-11.45
73	Cortisol[#]		**6.20**	5.75	6.55	8.43	6.40	8.74	5.89	-9.91	-11.18
74	Etiocholanolone[#]		**6.15**	6.96	7.04	7.07	5.99	7.74	6.03	-13.85	-12.12
75	Estradiol-3-benzoate		**5.94**	5.70	6.00	8.95	7.18	8.89	7.91	-12.56	-14.37
76	6-dehydro-19-Nortestosterone		**5.94**	5.84	6.13	8.36	5.79	8.39	5.45	-13.57	-3.73
77	4-Nonylphenol		**5.92**	5.39	6.94	6.82	5.91	8.52	5.80	-9.97	-8.52

#	Name	Structure	pK_d	I	II	III	IV	V	VI	VII	VIII
78	4-tert-Octylphenol		**5.67**	5.86	6.59	8.18	5.70	9.19	5.98	-8.95	-9.97
79	3,4-dicyclohexyl-hexane (meso)		**5.61**	7.62	7.95	8.58	5.91	7.74	6.00	-9.45	-10.27
80	Aldosterone[#]		**5.32**	5.35	5.14	5.34	5.58	5.34	5.12	-10.39	-8.33
81	17α-aminopropyl-17βb-hydroxy-5α-Androstan-3-one		**4.96**	5.66	5.50	7.22	4.54	6.82	5.05	-12.96	-11.98
82	Naringenin		**4.55**	4.71	4.28	7.70	4.73	7.45	4.63	-10.57	-10.69
83	3,4,-di [(4-hydroxyl)cyclohexyl] hexane (meso)		**4.54**	6.04	7.62	7.28	4.76	8.96	4.99	-11.51	-11.26
84	Genistein		**4.40**	4.72	4.98	8.40	4.50	7.36	4.16	-10.78	-11.04

SHBG LIGANDS IDENTIFIED IN THE CURRENT STUDY											
#	**Name**	**Structure**	***pK*$_d$**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**
85	ZINC00389056		**6.97**	8.01	5.18	8.52	6.09	7.63	6.97	-12.84	-11.59
86	ZINC00073647		**6.39**	6.95	5.42	9.24	10.2	8.90	8.19	-12.14	-12.76
87	ZINC00407192		**5.66**	5.80	6.14	8.08	7.76	8.38	7.67	-12.59	-12.59
88	ZINC02819939		**6.23**	5.80	5.56	9.16	10.2	9.06	7.34	-11.23	-11.59
89	ZINC00334865		**6.08**	5.76	6.60	8.37	4.56	8.24	4.40	-11.53	-12.14
90	ZINC00001785		**6.02**	5.88	6.15	8.25	4.38	7.83	3.83	-11.2	-10.08
91	ZINC00457465		**5.66**	8.97	3.56	8.21	7.24	8.47	7.73	-12.32	-12.41

*Compounds are presented along with the corresponding protein-ligand dissociation parameters pK_d, and predicted target affinities produced by QSAR and virtual screening approaches. Identification codes for compounds **85-91** correspond to internal IDs of the ZINC molecular database*

** these entries could not be scored in silico, as they failed to dock into 1LHN active site;
# 21 steroids forming the original 'steroid benchmark set';
$ we have used the corrected $pK_d = 7.44$ value for Deoxysortisol instead of $pK_d = 7.20$

The columns labeled with Roman numbers contain predictions by various QSAR models.
I – predictions by *QuaSAR-Evolution* model created on the basis of the updated dataset.
II GFA-QSAR model trained on the updated dataset (84 entries).
III - CoMFA model trained on the original dataset (21 entries).
IV - CoMFA model trained on the updated dataset (84 entries).
V - CoMSIA model trained on the original dataset (21 entries).
VI - CoMSIA model trained on the updated dataset (84 entries).
VII – results of 1LHN XP-docking with Glide.
VIII – results 1KDM XP-docking with Glide.

Table 2.2

Training and testing statistics for computational models created and their combinations investigated in the current study

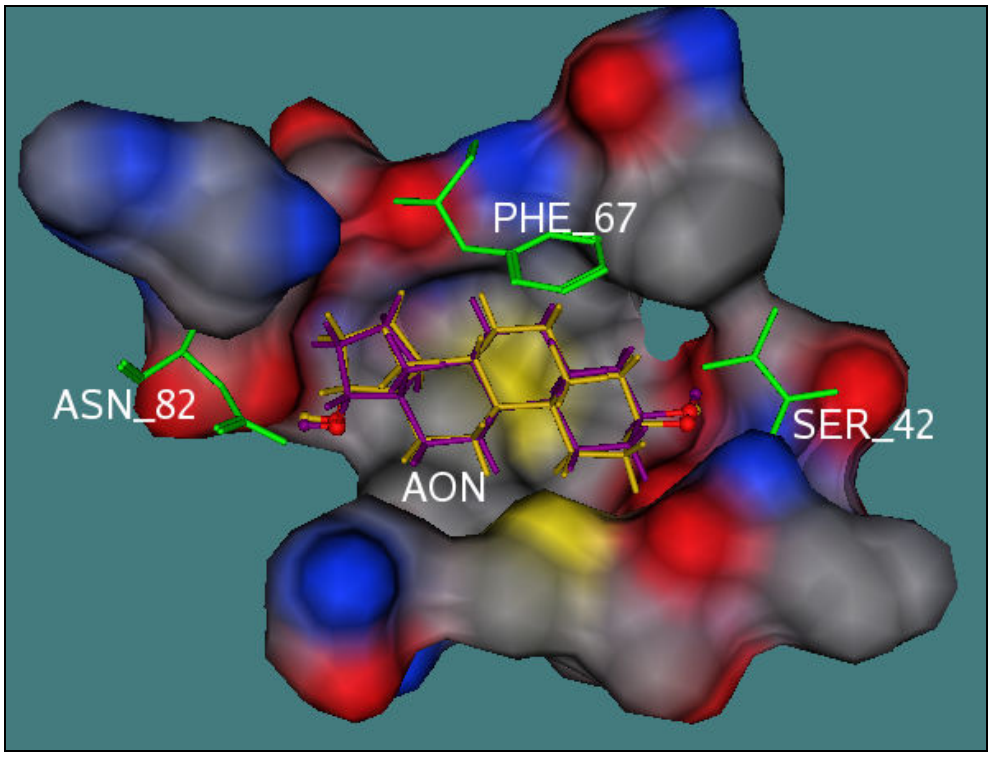
Models	r^2	q^2_{LOO}	*Training set* ***EF***
QSAR_GFA_7.0Å	0.56	0.44	3.5
QuaSAR_Evolution_7.0Å	0.66	0.58	3.5
CoMFA_corrected_21	0.99	0.45	3.0
CoMFA_traditional_21	0.98	0.53	2.0
CoMFA_84	0.99	0.41	5.5
CoMSIA_corrected_21	0.99	0.51	2.0
CoMSIA_traditional_21	0.98	0.53	1.0
CoMSIA_84*	0.91	0.49	5.0
1kdm_XP_GlideScore			2.5
1lhn_XP_GlideScore			2.0
2Dock_CoMFA84_CoMSIA84_GFA_QuaSAR			5.0
2Dock_CoMFA84_GFA			5.0
2Dock_CoMSIA84_GFA			4.0
2Dock_CoMFA84_CoMSIA84_GFA			5.0

**Three outliers were removed when training this model.*

The 'corrected' notion reflects alignment of SHBG ligands based on the docking poses. The 'traditional' notion corresponds to steroid scaffold alignments used in the original CoMFA and CoMSIA studies and based on maximal molecular field similarity.

Figure 2.1

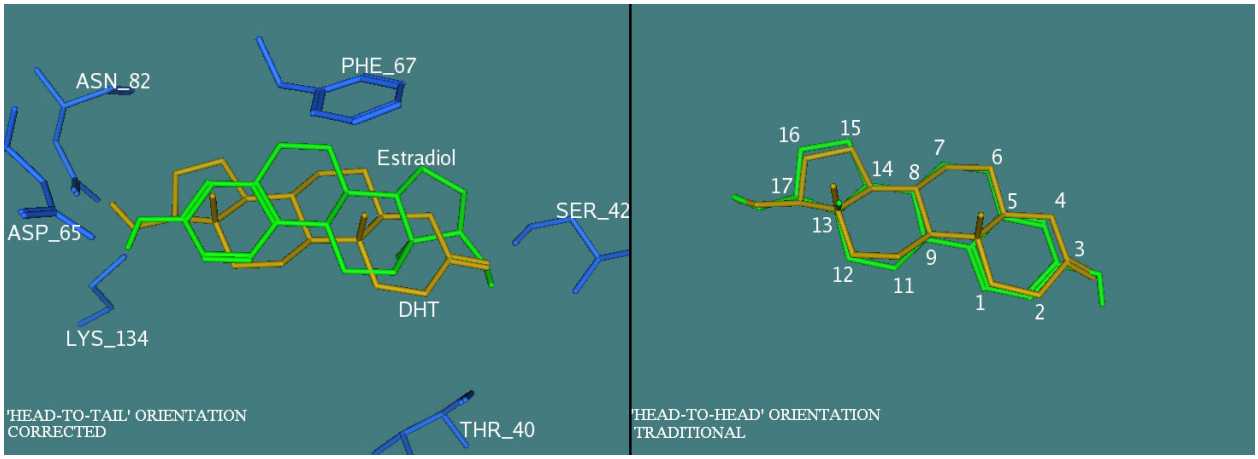
The co-crystallized ligand from 1LHN superposed with the docking pose of the same ligand



Superposition of the native ligand, 5α-androstane-3β,17α–diol (AON), from the 1LHN protein structure (colored in maroon) with the docking pose of AON established by the extra precision Glide protocol (in gold).

Figure 2.2

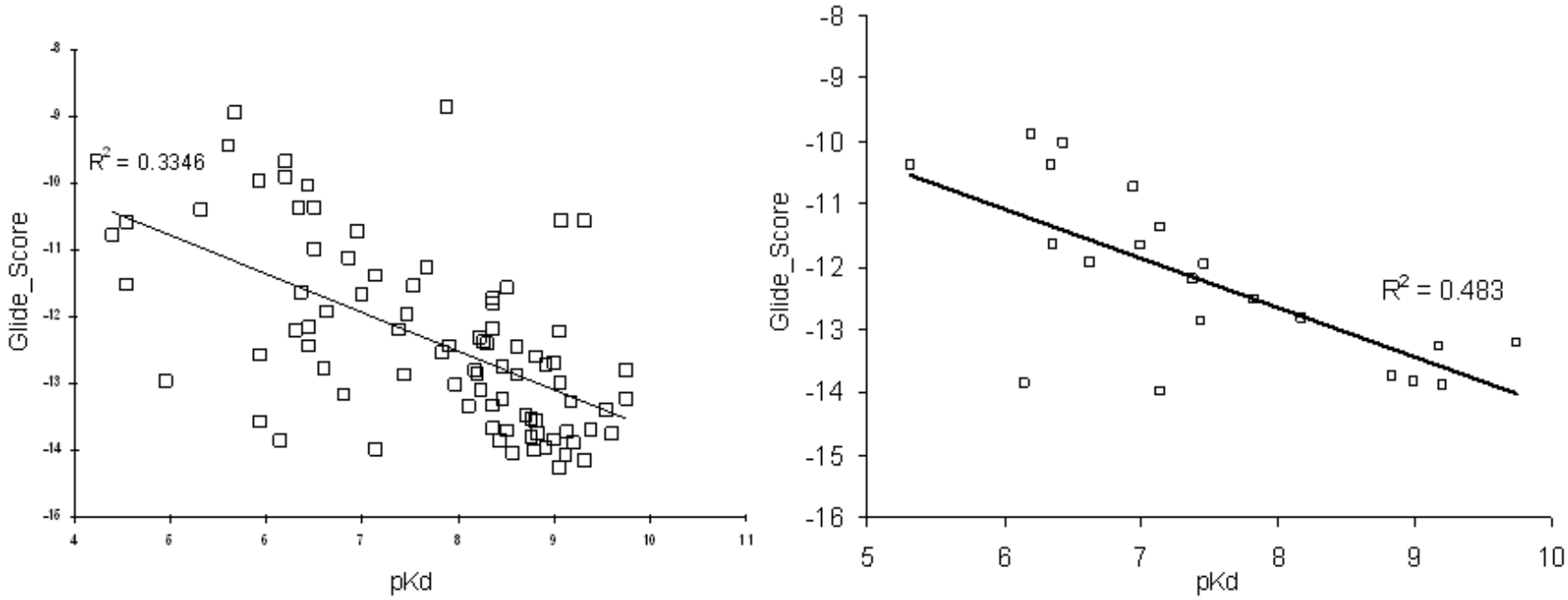
Optimal and traditional orientations of 5α-dihydrotestosterone (DHT) and estradiol in SHBG



Optimal (left panel) and traditional (right panel) orientations of 5α-dihydrotestosterone (DHT) is shown in gold, and estradiol shown in green. The correct superposition of the compounds within the SHBG steroid-binding site was derived from the 1KDM and 1LHU crystal structures. The traditional alignment was obtained by SYBYL.

Figure 2.3

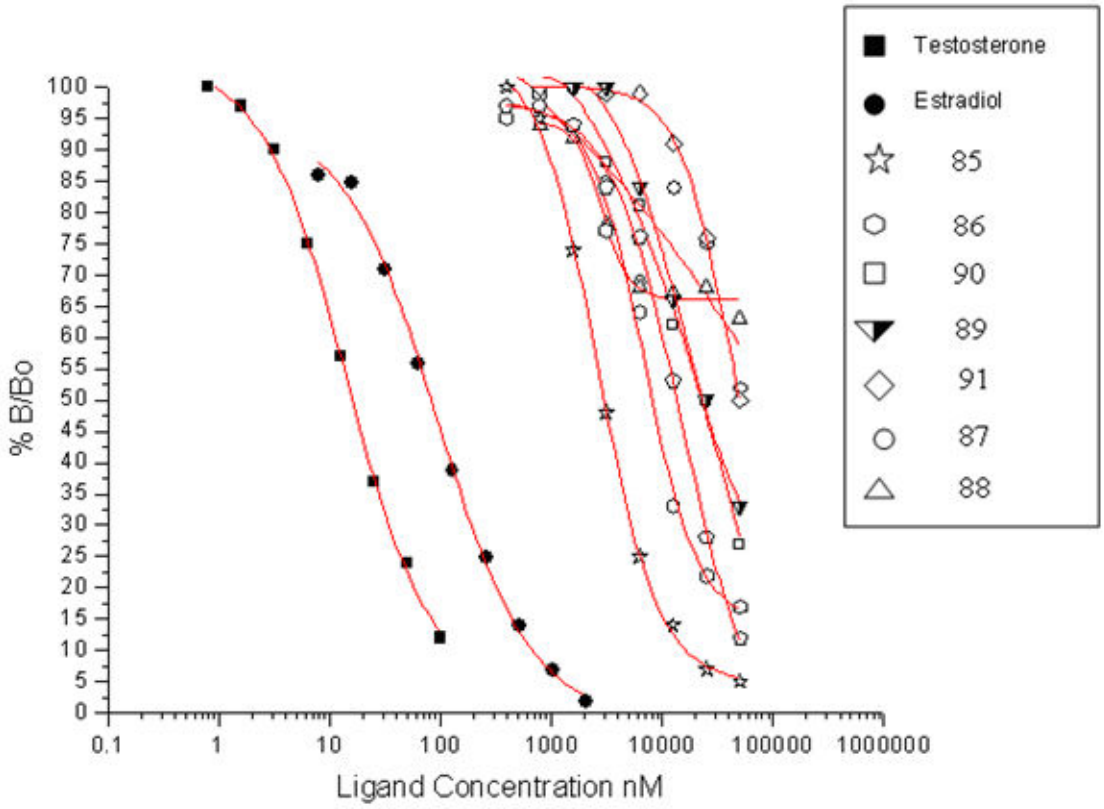
The linear dependence between Glide score values of and the corresponding experimental pK_d values



Left: the linear dependence between GlideScore values estimated by extra precision docking of 84 training set compounds and the corresponding experimental pK_d values (3 outliers have been removed) Right: the same dependence limited to 21 benchmark steroids

Figure 2.4

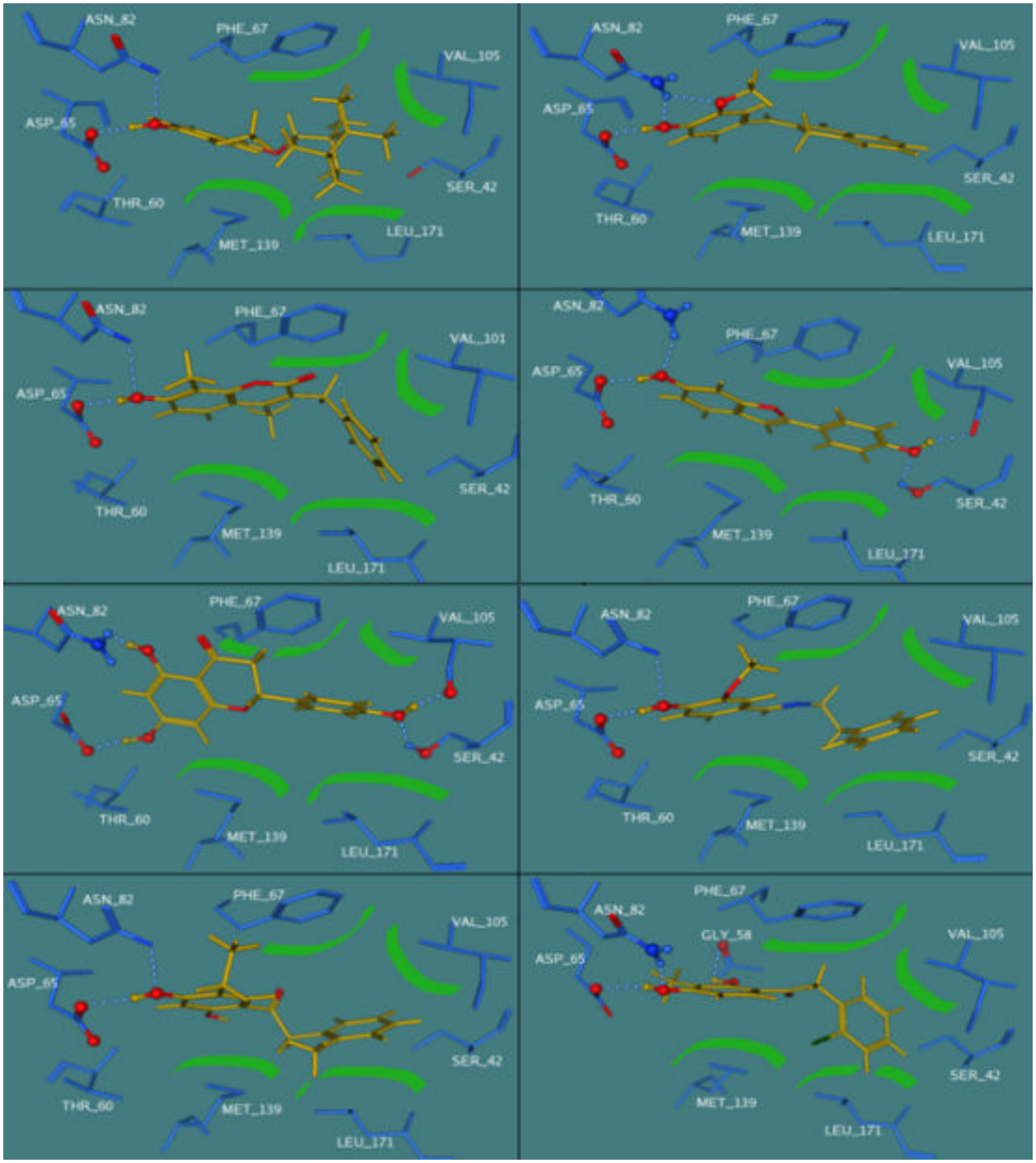
The displacement curves for test compounds used in the in vitro competition assay to determine the relative binding affinities of SHBG ligands



The displacement curves for test compounds used in the in vitro competition assay to determine the relative binding affinities of SHBG ligands. The percentage of tritium-labeled 5α-dihydrotestosterone bound to SHBG in the presence of increasing concentrations of competitor ligands.

Figure 2.5

Docked poses of the most active non-steroidal ligands within the hSHBG binding pocket



Only those residues that are most relevant to ligand binding are shown. Hydrogen bonds are represented as white dots; hydrophobic interactions featured by thick green lines. The following eight compounds are shown (ordered from left to right and top to bottom in the figure): **86,** ZINC00084751, **87, 89, 90, 88, 85** and **91**

2.6 References

Asikainen, A. H., Ruuskanen, J., and Tuppurainen, K. A. (2004). Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds. *SAR QSAR Environ. Res.* **15**, 19-32.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242.

Chem21 Group Inc. 2007. WOLF package, version 6.2

Chemical Computing Group, Inc. (2006). MOE: *Molecular Operating Environment*; Version 2006.08**,** Montreal, Canada

Chemical Computing Group, Inc. (2005). SVL exchange: http://svl.chemcomp.com/viewcat.php

Cherkasov A., Galkin V. I., Cherkasov R. A. (1996). The problem of the quantitative evaluation of the inductive effect: correlation analysis. Russian Chemical Reviews. **65**, 641-656.

Cherkasov A. 'Inductive' Descriptors. 10 Successful Years in QSAR. *Curt Comp-Aided Drug Design.* 1 (2005), pp. 21-42.

Cherkasov, A., Shi, Z., Fallahi, M., and Hammond, G. L. (2005). Successful in silico discovery of novel nonsteroidal ligands for human sex hormone binding globulin. *J. Med. Chem.* **48**, 3203-3213.

Cherkasov, A., Shi, Z., Li, Y., Jones, S. J., Fallahi, M., and Hammond, G. L. (2005). 'Inductive' charges on atoms in proteins: comparative docking with the extended steroid benchmark set and discovery of a novel SHBG ligand. *J. Chem. Inf. Model.* **45**, 1842-1853.

Cherkasov, A. (2006). Can 'Bacterial-Metabolite-Likeness' model improve odds of 'in silico' antibiotic discovery? *J. Chem. Inf. Model* **46**, 1214-1222.

Cherkasov, A., Ban, F., Li, Y., Fallahi, M., and Hammond, G. L. (2006). Progressive docking: a hybrid QSAR/docking approach for accelerating in silico high throughput screening. *J. Med. Chem.* **49**, 7466-7478.

Cramer, R. D., Patterson, D. E., and Bunce, J. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110**, 5959-5967.

Feher, M. (2006). Consensus scoring for protein-ligand interactions. *Drug Discov. Today* **11**, 421-428.

Friedman J. Multivariate Adaptive Regression Splines, Technical Report No. 102, Laboratory for Computational Statistics, Department of Statistics, Stanford University; Stanford, CA, Nov 1988 (revised Aug 1990).

Grishkovskaya, G. V., Avvakumov, G. V., Hammond, G. L., Catalano, M. G., Muller, Y. A. (2002a). Steroid ligands bind human sex hormone binding globulin in specific orientations and produce distinct changes in protein conformation. *J. Biol. Chem*. **207** 32086-32093.

Grishkovskaya, I., Avvakumov, G. V., Hammond, G. L., and Muller, Y. A. (2002a). Resolution of a disordered region at the entrance of the human sex hormone-binding globulin steroid-binding site. *J. Mol. Biol.* **318**, 621-626.

Halgren, T. A. (1996). Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization and Performance of MMFF94. *J. Comp. Chem*. **17** 490-519.

Hammond, G. L., Avvakumov, G. V., and Muller, Y. A. (2003). Structure/function analyses of human sex hormone-binding globulin: effects of zinc on steroid-binding specificity. *J. Steroid Biochem. Mol. Biol.* **85**, 195-200.

Hammond, G. L., and Lahteenmaki, P. L. (1983). A versatile method for the determination of serum cortisol binding globulin and sex hormone binding globulin binding capacities. *Clin. Chim. Acta.* **132**, 101-110.

Holland J. H. (1975). *Adaptation in Natural and Artificial Systems*, Ann Arbor, MI

Irwin, J. J., and Shoichet, B. K. (2005). ZINC -a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model* **45**, 177-182.

Jain, A. N., Koile, K., and Chapman, D. (1994). Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **37**, 2315-2327.

Karakoc, E., Cherkasov, A., and Sahinalp, S. C. (2006). Distance based algorithms for small biomolecule classification and structural similarity search. *Bioinformatics* **22**, e243-251.

Karakoc, E., Sahinalp, S. C., and Cherkasov, A. (2006). Comparative QSAR- and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. *J. Chem. Inf. Model* **46**, 2167-2182.

Klebe, G., Abraham, U., and Mietzner, T. (1994). Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **37**, 4130-4146.

Klebe G. (2006). Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today*. **11**, 580-594.

Korhonen, S. P., Tuppurainen, K., Laatikainen, R., and Perakyla, M. (2003). FLUFF-BALL, a template-based grid-independent superposition and QSAR technique: validation using a benchmark steroid data set. *J. Chem. Inf. Comput. Sci.* **43**, 1780-1793.

Liu, S. S., Yin, C. S., Li, Z. L., and Cai, S. X. (2001). QSAR study of steroid benchmark and dipeptides based on MEDV-13. *J. Chem. Inf. Comput. Sci.* **41**, 321-329.

Liu, S. S., Yin, C. S., and Wang, L. S. (2002). Combined MEDV-GA-MLR method for QSAR of three panels of steroids, dipeptides, and COX-2 inhibitors. *J. Chem. Inf. Comput. Sci.* **42**, 749-756.

Nyburg S. C. (1974). Some Uses of a Best Molecular Fit Routine. *Acta. Cryst.* **B30**, 251-253.

Polanski, J., and Walczak, B. (2000). The comparative molecular surface analysis (COMSA): a novel tool for molecular design. *Comput. Chem.* **24**, 615-625.

Robinson, D. D., Winn, P. J., Lyne, P. D., and Richards, W. G. (1999). Self-organizing molecular field analysis: a tool for structure-activity studies. *J. Med. Chem.* **42**, 573-583.

Rogers D. (1991). G/SPLINES: A Hybrid of Friedman's Multivariate Adaptive Regression Splines (MARS) Algorithm with Holland's Genetic Algorithm. *The Proceedings of the Fourth International Conference on Genetic Algorithms*, San Diego, July 1991.

Rogers D. (1992). Data Analysis using G/SPLINES. *Advances in Neural Processing Systems 4*; Kaufmann, San Mateo, CA.

Rogers D., Hopfinger A. J. (1994). Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **34** 854-866.

Schrödinger Inc. (2004). Maestro. San Diego, CA

Schrödinger Inc. (2006). Glide; Version 4.0, San Diego, CA.

Stahle, L., and Wold, S. (1988). Multivariate data analysis and experimental design in biomedical research. *Prog. Med. Chem.* **25**, 291-338.

Tripos, Inc. (2006). SYBYL, version 7.2. St. Louis, MO

Tuppurainen, K., Viisas, M., Laatikainen, R., and Perakyla, M. (2002). Evaluation of a novel electronic eigenvalue (EEVA) molecular descriptor for QSAR/QSPR studies: validation using a benchmark steroid data set. *J. Chem. Inf. Comput. Sci.* **42**, 607-613.

Tuppurainen, K., Viisas, M., Perakyla, M., and Laatikainen, R. (2004). Ligand

intramolecular motions in ligand-protein interaction: ALPHA, a novel dynamic descriptor and a QSAR study with extended steroid benchmark dataset. *J. Comput. Aided Mol. Des.* **18**, 175-187.

Turner, D. B., Willett, P., Ferguson, A. M., and Heritage, T. W. (1999). Evaluation of a novel molecular vibration-based descriptor (EVA) for QSAR studies: 2. Model validation using a benchmark steroid dataset. *J. Comput. Aided Mol. Des.* **13**, 271-296.

Warren, G. L., Andrews, C. W., Capelli, A. M., Clarke, B., LaLonde, J., Lambert, M. H., Lindvall, M., Nevins, N., Semus, S. F., Senger, S., Tedesco, G., Wall, I. D., Woolven, J. M., Peishoff, C. E., and Head, M. S. (2006). A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **49**, 5912-5931.

Westphal, J. (1971). Steroid-protein interactions. Monographs on endocrinology, Vol. 4. Springer 567

3 In Silico Identification of Anthropogenic Chemicals as Ligands of Zebrafish Sex Hormone Binding Globulin*

3.1 List of Authors

Nels Thorsteinson, Fuqiang Ban, Osvaldo Santos-Filho, Seyed M.H. Tabaei, Solange Miguel-Queralt, Caroline Underhill, Artem Cherkasov and Geoffrey L. Hammond

3.2 Introduction

As in other vertebrates, sex hormone binding globulin (SHBG) transports sex steroids in the blood of fish and regulates their access to tissues (Siiteri et al., 1982). In addition to binding its natural steroid ligands, SHBG in humans bind xenobiotics including potential endocrine disruptors present in waste water systems (Hodgert-Jury et al. 2000). Since aquatic species are in intimate contact with these anthropogenic

---

* A version of this chapter, co-authored by Thorsteinson, N., Ban, F., Santos-Filho, O., Tabaei, S.M.H, Miguel-Queralt, S., Underhill, C., Cherkasov, A. and Hammond, G.L., has been submitted for publication to *Toxicology and Applied Pharmacology*

Abbreviations: CAS, chemical abstract service; CEPA, Canadian environmental protection act; CoMFA, comparative molecular field analysis; CoMSIA, Comparative molecular similarity index analysis; DBP, 9,10-dihydrobenzo(a)pyren-7(8H)-one; DHT, 5α-dihydrotestosterone; DSL, domestic substances list; EE, ethinylestradiol; EINECS, European inventory of existing chemical substances; GS: Glide Score; hSHBG, human sex hormone binding globulin; LIE, linear interaction energy; LOO, leave one out; MD, molecular dynamics; MMFF Merck molecular forcefield; MOE, molecular operating environment; NDSL, non-domestic substances list; OC, 4-tert-octylcatechol; OP, 4-tert-octylphenol; PAH, polycyclic aromatic hydrocarbon; PDB, protein data bank; PLS, partial least squares; RBA, relative binding affinity; SD format, structure data format; SHBG, sex hormone binding globulin; zfSHBG, zebrafish sex hormone binding globulin; SMILES, simplified molecular input line entry specification; SVL, scientific vector language

compounds, fish SHBGs represent interesting targets for computational toxicology studies of potentially harmful environmental contaminants.

Human SHBG (hSHBG) along with the steroid benchmark set of ligands has been a model system for computer-aided drug design research (Tuppurainen et al., 2004; Asikainen et al., 2004; Korhonen et al., 2003; Liu et al., 2002; Tuppurainen et al., 2002; Liu et al. 2001; Cherkasov et al., 2005; Cherkasov et al., 2008; Cramer et al., 1988; Klebe et al., 1994). The binding data of the steroid benchmark set has been used to develop popular molecular modeling tools such as comparative molecular field analysis (CoMFA) (Cramer et al., 1988) and molecular similarity indices in a comparative analysis (CoMSIA) (Klebe et al., 1994). Recently, novel non-steroidal nanomolar ligands of hSHBG have been identified by applying such methods to an alignment-corrected version of the benchmark steroids (Cherkasov et al., 2008).

The biological importance of SHBG in fish is not as well studied as in mammals. It has been demonstrated that the protein is expressed primarily in the liver and intestine of zebrafish (*Danio rerio*) (Miguel-Queralt et al., 2004), and the uptake of steroids from their aquatic environment appears to be influenced in some way by their affinity for SHBG (Scott et al., 2005). Given the role SHBG may play in regulating the bioavailability of androgens and estrogens during sexual differentiation and the reproductive cycle in fish (Miguel-Queralt et al., 2007), environmental compounds that bind to fish SHBG could adversely influence their reproductive performance. Ethinylestradiol (EE) was of particular interest in this context because it is a well-known endocrine disruptor in some fish and has an unusually high affinity for SHBG in zebrafish (Miguel-Queralt and Hammond, submitted). Moreover, synthetic ligands of fish SHBG sequestered from water may accumulate in the bodies of the fish and

subsequently harm predators, including humans. Public concern about the potential toxicity of such xenobiotic substances in the oceans, lakes, and rivers has prompted several government funded environmental agencies, such as the European Chemicals Bureau and Environment Canada, to identify commercial substances that represent health or environmental risks.

The sequence of zebrafish SHBG (zfSHBG) has been reported together with values of its relative binding affinity (RBA) for 19 steroids from the steroid benchmark set (Miguel-Queralt et al., 2004). In the current work, a zfSHBG homology model was built from an hSHBG crystal structure template (Grishkovskaya et al., 2002b). Molecular dynamics (MD) simulations were performed to refine the model; to explain previously observed zfSHBG steroid-binding characteristics; to identify amino acid substitutions responsible for the unique ligand binding properties of hSHBG and zfSHBG, and to assess the accuracy of the model.

The zfSHBG model was then used in a multi-method virtual screening pipeline similar to a previously developed strategy (Cherkasov et al., 2008) which involved large-scale docking and CoMFA, and CoMSIA modeling to identify potential anthropogenic zfSHBG ligands from the large ZINC chemical database (Irwin and Shoichet, 2005) and from lists of existing commercial chemicals, some of which were subsequently validated experimentally in a zfSHBG ligand-binding assay.

3.3 Materials and Methods

3.3.1 ZINC chemical database preparation

Almost four million structures in SD format from the ZINC database were imported into a database using Molecular Operating Environment (MOE) version 2006.08 (Chemical Computing Goup Inc., 2006). These structures were washed –*i.e.* all

inorganic components were removed, and all ionizable groups were coordinated with pH=7.0 conditions. Next, the database was energy minimized using the MMFF94x (Halgren, 1996) forcefield and exported in SD format for use by the Glide (Schrödinger Inc, 2006) docking program.

3.3.2 Commercial chemical database preparation

Structures for over 80 000 compounds from various environmental compounds lists were obtained. The lists include the European inventory of existing commercial substances (EINECS) (European Chemicals Bureau, 2002) and the Canadian Environmental Protection Act (CEPA) environmental registry's domestic substances list and non-domestic substances list (DSL and NDSL) (Environment Canada, 2006). The EINECS contains substances manufactured or imported into the European Union and has a high overlap with CEPA's DSL and NDSL. The 80 000 compounds obtained were those within the grasp of our resources and they represent the majority of the substances in the lists.

We obtained 68 970 substances in simplified molecular input line entry specification (SMILES) format from the European Chemicals Bureau (http://ecb.jrc.it/qsar/information-sources/). Most of the remaining EINECS and CEPA structures were also obtained from various online resources as SMILES strings, but we were not able to obtain the remaining EINECS and CEPA substances. In the end, SMILES strings for approximately 80 000 compounds, representing about 70 % of the EINECS and CEPA lists, were obtained for in silico screening of zfSHBG binding.

The 80 000 SMILES strings were imported into a MOE database. MOE was used to rebuild the SMILES strings into 3D structures. These structures were washed –*i.e.* all inorganic components were removed, and all ionizable groups were coordinated with

pH=7.0 conditions. Next, the database was energy minimized using the MMFF94x forcefield and exported in SD format for use by the Glide docking program.

3.3.3 Homology modeling of zfSHBG

SHBG protein sequences from five different species: zebrafish, rainbow trout, European seabass, mouse, and human were obtained from the NCBI protein database and correspond to accession numbers AAU14174, BAE48779, AAW23033, NP_035497, and AAC18778, respectively. A multiple alignment of these five sequences was performed and the hSHBG to zfSHBG amino acid mapping within this alignment was used for homology modeling. MOE was utilized with default settings to construct the zfSHBG homology model from the hSHBG template structure of the 1KDM entry of the protein data bank (PDB).

3.3.4 Molecular dynamics and binding free energy calculations

All MD simulations and binding free energy calculations were performed on zfSHBG using the GROMACS 3.3 simulation package (van der Spoel et al., 2005). Newton's equations of motion were integrated with a time step of 1.5 fs. Short-range and long-range forces were cut-off at 0.9 nm and 1.4 nm respectively and periodic boundary conditions were applied. The simulations were conducted at 300 K.

The linear interaction energy (LIE) method (Hansson et al., 1998) was used to calculate the binding free energies (ΔG_{bind}). For each ΔG_{bind} calculation, two 750 ps NVT simulations (constant moles, volume, and temperature) were performed; one with the protein bound with the ligand in water, and one with just the ligand in water. PRODRG (Schuttelkopf and van Aalten, 2004) was used to generate the ligand topologies according to the GROMACS87 forcefield but the ligand Gasteiger (Gasteiger and Marsili, 1980) partial charges were calculated using MOE. The Coulomb and

Lennard-Jones interaction between ligand and solvent from the 750 ps simulations were used in the LIE formula with the following parameters: β=0.33, γ=0, α=0.18.

3.3.5 Molecular docking

The Maestro suite (Schrödinger Inc., 2004) was used to prepare the 1LHN hSHBG structure and the zfSHBG model for docking. All water and ion molecules were removed from the corresponding PDB files, and hydrogen atoms were added and adjusted where necessary. The steroid-binding sites were defined as 10 Å surrounding the ligands in all cases. Docking was performed using Glide 4.0 parallel suite with default settings.

The MOE estimated pK_i was calculated for each ligand using the scoring.svl script available through the SVL exchange service (Chemical Computing Group, Inc., 2005). For this, hydrogen atoms were added to the zfSHBG model or to 1LHN and the partial charges were calculated with the AMBER99 forcefield (Wang et al., 2000). The Gasteiger partial charges were calculated for the structures that passed the docking cut-off. The estimated pK_i for these structures were calculated by choosing the dock_pKi descriptor with default settings for the molecular database.

3.3.6 CoMFA modeling

A set of 19 ligands from the steroid benchmark set with experimental pK_as in zfSHBG were docked using Glide. The resulting docking poses were used to build CoMFA and CoMSIA models. For hSHBG, 87 steroids from (Cherkasov et al., 2008) were used to build the models. Any ligands docked incorrectly were manually repositioned into the correct orientation according to crystallographic information (Grishkovskaya et al., 2002a).

The zfSHBG and hSHBG ligands were used to compute steric and electrostatic CoMFA fields. The steric fields were calculated on 2 Å grids, by evaluating '6-12' van der Waal's interactions with default CoMFA probes. We used the distance-dependent dielectric parameters to compute the Coulomb interactions and approximated electrostatic CoMFA fields with the truncation parameter set to 30.0 Kcal/mol. The SYBYL package (Tripos Inc., 2004) was used to construct all CoMFA models using the partial least squares (PLS) fitting. Cross-validation was carried out by the built-in LOO procedure.

3.3.7 CoMSIA modeling

Using the experimental zfSHBG and hSHBG ligands, we computed five CoMSIA properties that included steric, electrostatic, hydrophobic, hydrogen bond donor, and hydrogen bond acceptor fields (computed with default settings). The fields were derived according to the similarity indices of molecules brought into a common alignment using an attenuation factor of 0.3. Each CoMSIA property of a given atom was scaled to 74.1 % for its 1 Å proximity, to 30.1 % for >2 Å surrounding, and to 6.7 % for the area within 3 Å.

The final CoMSIA models were constructed using the PLS algorithm and cross-validated by the LOO procedure implemented by the SYBYL package.

3.3.8 Expression of recombinant zfSHBG and site-directed mutagenesis

Recombinant zfSHBG was stably expressed in Chinese hamster ovary cells using a full-length zfSHBG cDNA cloned into the pRc/CMV expression vector, as previously described (Miguel-Queralt et al., 2004). After selection in the presence of Geneticin (Invitrogen), stably transfected cells were grown to near confluence, washed twice with phosphate-buffered saline to remove fetal bovine serum, and then cultured in

HyQ PF-CHO LS (HyClone, Logan, UT) media for 3-5 days. The medium was then harvested and stored at 4 ºC until used for ligand-binding studies (see below).

Site-directed mutagenesis of the zfSHBG cDNA within the expression vector was performed using the QuikChange® XL Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA) and the following mutagenic oligonucleotide: 5' CCTGAAATGCAGATTGGAATGCAGACATCTTAGTGAG in which the altered nucleotide (underlined) converted a Lys codon (AAG) into a Asn codon (AAT). The mutated cDNA was sequenced to confirm that only the targeted mutation had occurred. The resulting mutant and wild-type zebrafish SHBG cDNAs were then expressed in CHO cells, as described above.

3.3.9 Ligand-binding assays

An established competitive ligand-binding assay was used to determine the RBAs of the test compounds to zfSHBG (Miguel-Queralt et al., 2004), when compared to 5α-dihydrotestosterone (DHT). In brief, 100 µl recombinant zfSHBG were incubated for at least 1h at room temperature with tritium labeled DHT ([^{3}H] DHT) in the presence or absence of excess unlabeled DHT to monitor non-specific binding. After a further 1 h incubation at 0 ºC, 500 µl of a dextran-coated charcoal slurry was added at 0 ºC, and incubated for 8 min prior to centrifugation to separate zfSHBG-bound from free [^{3}H] DHT. For the screening assay, triplicate aliquots of fixed amounts (33 -50 µM) of test compound were assayed. Compounds that displaced more than 95 % of the [^{3}H] DHT from the zfSHBG in this assay were then diluted serially to generate complete competition curves. The amounts of [^{3}H] DHT bound to zfSHBG at each concentration of competitor ligand were measured and plotted in relation to the amount of [^{3}H] DHT bound to zfSHBG at zero concentration of competitor. From the resulting competition

curves, IC_{50} concentrations were calculated if displacement of more than 50 % of [^{3}H] DHT from zfSHBG was achieved.

The dissociation constants (K_d) have been calculated from the relative binding affinity parameters using the following equation: $1/\{K_a(DHT)/[(1+R)/RBA - R]\}$, where $K_a(DHT) = 0.98 * 10^9\ M^{-1}$ is the association constant of the DHT and R (0.05) is the ratio of bound-to-free [^{3}H] DHT at 50 % displacement in the assay.

3.4 Results

3.4.1 Homology modeling of zfSHBG

MOE was used with default settings to construct the zfSHBG homology model from the 1KDM hSHBG template structure and using the sequence alignment illustrated in Figure 3.1. Although this is not the highest resolution structure of hSHBG, the 1KDM template was used because the loop region that covers the entrance to the hSHBG steroid-binding site is ordered and visible in this 2.35 Å structure (Grishkovskaya et al., 2002b).

3.4.2 Visual inspection of key contact points between ligands and amino acid residues within the hSHBG and zfSHBG steroid-biding sites

Here we examine the interactions that likely occur when testosterone, estradiol, DHT or androstenedione are positioned within the hSHBG and zfSHBG active sites. In these comparisons, the amino acids within hSHBG and zfSHBG are numbered according to their positions in their respective mature polypeptide sequences (Petra et al., 1986; Miguel-Queralt et al., 2004). For reference, the experimental pK_a values of natural steroids, as well as for EE, are shown in table 1 in the contexts of human and zebrafish SHBGs.

Testosterone, estradiol, and DHT each form three hydrogen bonds with the Ser42, Asp65, and Asn82 residues of hSHBG (Figure 3.2). These hydrogen bonds, along with hydrophobic contacts from nearby hydrophobic residues such as Phe67 explain why testosterone, estradiol, and DHT all bind to hSHBG with high affinity. Androstenedione can only hydrogen bond with Ser42, although a weak hydrogen bond with Asn82 is possible, and this accounts for its poor affinity for the hSHBG steroid-binding site. By contrast, androstenedione has very high affinity for zfSHBG, and this can be explained by the Lys76 in the zfSHBG active site that would allow for a strong hydrogen bond with the oxygen at C17 of androstenedione.

3.4.3 Identification of amino acids contributing to differences in the binding of ethinylestradiol to hSHBG and zfSHBG

Experimental analyses (Miguel-Queralt et al., 2004; Hodgert-Jury et al., 2000) have shown that EE has a much higher affinity for zfSHBG (pK_a = -9.72) than for hSHBG (pK_a = -6.81). We therefore used binding free energy calculations from MD simulations to determine which amino acid residues within hSHBG and zfSHBG are responsible for this difference. As a proof of principle, we first assessed how well the model might predict the experimental results of a study of estradiol binding to several hSHBG mutants (Hammond et al., 2003). The mutants were constructed in the context of the 1KDM hSHBG crystal structure using MOE. We used the LIE method to estimate the change in free energy as a result of ligand binding (ΔG_{bind}) from 750 ps MD simulations performed by the GROMACS 3.2 simulation program using the GROMACS87 force field. In accordance with the experimental values (Hammond et al., 2003), the model predicted that estradiol has a weaker binding affinity for the D65A, S42A, and G58A mutants than for wild type hSHBG or its W84A mutant (Table 3.2).

We then performed MD and ΔG_{bind} calculations on several novel in silico hSHBG mutants, each containing a substitution in the steroid-binding site corresponding to the amino acid in that position within zfSHBG. For each of these hSHBG mutants, we calculated the ΔG_{bind} of EE assuming EE is in the orientation where the C3 hydroxyl interacts with Asp65, i.e., the same orientation as estradiol in hSHBG observed by X-ray crystallography (Grishkovskaya et al., 2002a). In these models, EE bound to the N82K and V105L hSHBG mutants with lower ΔG_{bind} values than that for EE binding to wild type hSHBG (Table 3.3). These findings provide a possible explanation for why EE has a much greater affinity for zfSHBG than for hSHBG. In essence, they suggest that the zfSHBG Lys76 and Leu99 residues, which correspond to Asn82 and Val105 in hSHBG, contribute to the strong binding of EE to zfSHBG. The contribution of Lys76 in zfSHBG to its remarkably high affinities for EE and androstenedione (see above) when compared to hSHBG, was tested by producing a zfSHBG mutant in which Lys76 was substituted with an Asn, as found in the corresponding position of hSHBG. Although the steroid-binding affinity of the K76N zfSHBG mutant for DHT is reduced substantially when compared to the wild-type zfSHBG, the steroid-binding specificity of this mutant revealed 15 and 9 fold reductions in the relative binding affinities (RBAs) for androstenedione and EE, respectively, when DHT was used as the reference ligand in a competitive steroid-binding assay (Miguel-Queralt et al., 2004).

3.4.4 Binding free energies of ethinylestradiol bound to zfSHBG

To further validate our model, we performed a binding free energy analysis of the zfSHBG steroid-binding site. To do this, we again first calculated ΔG_{bind} of EE in each of its four possible binding orientations in the zfSHBG model. The results indicate that the orientation in which the C3 hydroxyl of EE hydrogen bonds to zfSHBG Asp59 (pose 1)

provides the strongest binding free energy with a predicted value of -26.30 KJ/mol (Figure 3.3). When EE is oriented so that the C17 hydroxyl hydrogen bonds to Asp59 (Figure 3.3, pose 2), it also had a relatively strong binding free energy of -25.03 KJ/mol, while the ΔG_{bind} of the two other possible orientations of EE in the binding site were only -21.25 KJ/mol and -20.64 KJ/mol and were not considered further.

As mentioned above, binding free energy calculations predicted strong binding for pose 2 (Figure 3.3). From looking at the zfSHBG active site optimized with EE bound in this orientation, several factors could contribute to this. The backbone oxygens of zfSHBG Ser36 and Leu99 could form hydrogen bonds with the hydroxyl group at C3 of the aromatic ring of EE, and Asp59 and Lys76 could hydrogen bond with the hydroxyl group at the C17 of its 5-carbon ring. However, as mentioned above and in line with the slightly higher calculated binding free energy value for pose 1, it is more likely that EE resides within the zfSHBG binding pocket in the same orientation as estradiol in hSHBG crystal structures. In this orientation, the functional moieties of EE would hydrogen bond with Lys76 (hydroxyl at C3), Asn59 (hydroxyl at C3), and Ser36 (hydroxyl at C17) within the zfSHBG ligand-binding pocket, and this would allow the large ethinyl group at C17 to be neatly packed forming hydrophobic contacts with Leu99 (Figure 3.3).

3.4.5 Docking of the benchmark steroids into zfSHBG

Docking experiments were performed to further validate the zfSHBG homology model. First, the zfSHBG model was refined by 7.5 ns MD simulation with testosterone in the steroid-binding site. The equilibrium zfSHBG structure achieved after approximately 5.4 ns of the 7.5 ns MD simulation was then used in all further analysis.

Glide docking scores were obtained for the 19 steroids of the benchmark set with pK_a values for zfSHBG calculated from the available ligand-binding data analyses

(Miguel-Queralt et al., 2004). The Pearson correlation squared r^2 between the predicted Glide scores and experimental pK_a values for these compounds is 0.28 (Figure 3.4). This r^2 value is sufficient for validating the accuracy of the homology model because we performed the same experiment using an hSHBG crystal structure (1LHN) and a set of 87 experimentally tested hSHBG ligands and this resulted in a similar r^2 of 0.26 as shown in Figure 3.4.

The results of the MD binding free energy experiments, and the correlation of Glide score to experimental pK_a for the 19 steroids tested, provide independent confirmations that the computational methods applied to the zfSHBG homology model produce reliable results.

3.4.6 Virtual screening the ZINC database for zfSHBG binders

We employed a multi-method screening pipeline similar to that successfully implemented by Cherkasov et al., 2008. About four million compounds from the ZINC database were docked to the zfSHBG model using the Glide docking program (see *Materials and Methods*). The four million docked ligands were narrowed down to 22 555 by applying a cut-off of -9.0 to the Glide docking scores.

Each of the remaining 22 555 compounds was then assigned predicted pK_a values based on both CoMFA and CoMSIA models. To create these models, the 19 original ligands with experimental pK_as for zfSHBG were docked using Glide, and their docking poses were obtained. Any ligands docked incorrectly were manually repositioned into the correct orientation according to the "alignment-corrected" steroid benchmark set (Cherkasov et al., 2008). The 19 docked zfSHBG ligands were used to compute steric and electrostatic CoMFA fields. We computed five CoMSIA properties: steric, electrostatic, hydrophobic, hydrogen bond donor, and hydrogen bond acceptor

fields. These CoMFA and CoMSIA models were used to calculate the estimated pK_as, making up two of the four predictors for the 22 555 ligands.

The fourth and final predictor calculated for each ligand was the 'dock_pKi' parameter computable by the MOE scoring.svl script. The dock_pKi is an estimated pK_i value defined by the energy of hydrogen bonds, transition metal interactions, and hydrophobic interactions. However, in the case of the zfSHBG active site, the dock_pKi is predominantly influenced by hydrophobic contacts. Thus, it is a score that is representative of the energy of favorable hydrophobic contacts between ligands and their receptors.

Based on the four predictors, we employed a community voting scheme which assigned a 1 vote to the top 10 % of each score while all other docked ligands were given a vote value of 0. Then, the four votes were added together for each ligand and we selected the 582 ligands with a score of 3 or 4. As expected, steroid derivatives were overly represented in the list of potential ligands, providing strong evidence that our in silico methods were effective in identifying zfSHBG ligands (the 582 structures in SD format will be available in the journal's online supplementary material).

3.4.7 Experimental testing of 42 compounds from the ZINC database for zfSHBG binding

The 42 compounds tested were originally selected from the ZINC database by a previous in silico study on hSHBG (Cherkasov et al., 2008) and we expected that some of these compounds might bind to the zebrafish homologue. In the screening assay, 25 of these in silico hSHBG hits demonstrated zfSHBG-binding competition with [^{3}H]DHT as the labeled ligand. Six of these compounds were able to displace more than 50 % of bound [^{3}H]DHT from zfSHBG. These six compounds (*ZINC00392254, ZINC00233302,*

ZINC00084751, ZINC00073647, ZINC00334865, ZINC02067917) did not cause any solubility concerns, and were further analyzed to generate complete competitive displacement curves (Figure 3.5). Their corresponding zfSHBG association constants (K_a) calculated from their substitution curves are between 1.0 x 10^6 M^{-1} and 1.0 x 10^8 M^{-1}.

The 22 555 compounds that made the Glide score cutoff of -9.0 contained 12 of the 42 compounds that were tested experimentally for zfSHBG binding. Four of these twelve substances belong to the set of six micromolar to nanomolar zfSHBG ligands identified. Since four out of twelve compounds that made the docking cutoff turned out to be high-affinity zfSHBG ligands, and because the lists of zfSHBG binders selected by virtual screening are dominated by steroids, we expect that it contains many other potential zfSHBG binders.

The six experimentally verified zfSHBG ligands are shown docked to the steroid binding site of the zfSHBG model in Figure 3.6. Each of these compounds appears to form a hydrogen bonds with Asp59 and/or Lys76 in zfSHBG and they all possess a hydrophobicity similar to steroids. Five of them lack an electronegative atom on the opposite side for hydrogen bonding with Ser36, and this is likely the reason why they bind with less affinity than steroidal ligands that can hydrogen bond with this residue. The exception, ZINC00334865, is highly aromatic and is likely to be unable to form the hydrophobic contacts necessary for optimal binding.

The contribution of interactions from Lys76 of zfSHBG to the binding of these six ligands was tested in a competitive binding assay using the K76N zfSHBG mutant described above. The binding of ZINC0023302, ZINC00084751, ZINC00073647, and ZINC00334865 to the K76N mutant is slightly reduced from that to wildtype, indicating

that Lys76 contributes in part to the binding of these compounds to zfSHBG and that Asp59 likely provides the more important hydrogen bond anchoring, as suggested by the binding orientation predictions of the Glide docking program. Although the Glide docking poses indicate that ZINC02067917 and ZINC00392254 form hydrogen bonds with Lys76, the binding of ZINC02067917 to the zfSHBG K76N mutant is not very different from its binding to wild-type zfSHBG, while ZINC00392254 actually binds zfSHBG K76N at lower concentrations than to wild-type zfSHBG. There are two possible reasons for this contradiction: either the Glide should have predicted a stronger interaction of these two compounds with Asp65 and little or no interaction with Lys76, or the Asn in the zfSHBG K76N mutant somehow compensates for or enhances their interactions with Lys76.

3.4.8 Virtual screening of the commercial substances lists

We were able to obtain SMILES strings for over 80 000 compounds, representing over 80 percent of the lists of existing commercial substances from the European Chemicals Bureau and Environment Canada (see *Materials and Methods* for more details on the commercial database preparation). MOE was used to convert the 80 000 SMILES strings into a database of energy minimized molecular structures and our in silico screening procedure was used to search this database for commercial compounds that bind to zfSHBG. In this case, we calculated the same four scores as in the ZINC database screening procedure for both hSHBG and zfSHBG, providing eight scores in total. We utilized both hSHBG and zfSHBG models because the hSHBG homologue binds many of the same ligands as zfSHBG, and we assumed that the hSHBG crystal structure would also be useful for scoring ligands for zfSHBG binding.

The first step was to calculate the Glide scores for each compound and to apply a cutoff for any compound with a Glide score less than -8.75 in either hSHBG or zfSHBG, or a Glide score sum (hSHBG + zfSHBG) less than -17.0. This initial filter removed all but 1 034 compounds from the original set of over 80 000. These two Glide scores (one for hSHBG and the other for zfSHBG) comprised two of the eight scores of the voting scheme.

The remaining 1 034 compounds were then each assigned predicted pK_a values based on both CoMFA and CoMSIA models. For zfSHBG, the same CoMFA and CoMSIA models described above were used. For hSHBG, the same procedure as described above for obtaining the CoMFA and CoMSIA models for zfSHBG was performed using a set of 87 steroids with experimental binding constants for hSHBG (Cherkasov et al., 2008). These models resulted in a total of four more predictors for the voting scheme: CoMFA and CoMSIA estimated pK_i for both zfSHBG and hSHBG.

The dock_pKi parameter (described in the ZINC database virtual screening section) for each of the 1 034 ligands was calculated for both hSHBG and zfSHBG, providing the final two predictors of the set of eight to be used for consensus voting.

Based on the eight scores, we implemented a voting system where, for each score, a value of 1 was assigned to the ligands in the top 35 % while all other ligands were given a value of 0. Then, the eight binary votes (Glide score, CoMFA, CoMSIA, and dock_pKi for both hSHBG and zfSHBG) were added together for each ligand and, in turn, the ligands were ranked by their voting score where a score of eight was the maximum and zero was the minimum. Of the 1034 ligands, 400 with the highest vote total were further examined by eye to assign priority to the ligands based on our experience with steroid binding interactions. After final visual inspection, 14 compounds

were distinguished as ‘top hits’, 6 of which were selected for experimental testing (Table 3.4).

3.4.9 Experimental validation of six in silico hits from the commercial substances set

The initial experimental testing of the six in silico hits was conducted at a single concentration (33 µM), and all of them i.e., hexestrol, DL-thyronine, 4-tert-octylcatechol (OC) and 9,10-dihydrobenzo(a)pyren-7(8H)-one (DBP), coumestrol and podocarpic acid, resulted in greater than 30% displacement of the [^{3}H]DHT from zfSHBG. These six compounds were therefore further analyzed as zfSHBG ligands in a concentration-dependent manner. The resulting competitive displacement curves (Figure 3.7) demonstrate that coumestrol, DL-thyronine and podocarpic acid show only limited binding affinity for zfSHBG, which precluded the measurement of an IC_{50} value required to determine their RBA values. By contrast, hexestrol, OC and DBP all bind with 2-3 orders of magnitude less affinity than the reference steroid (DHT), which would place them within the micromolar range. Thus, our virtual screening procedure effectively enriched for potential zfSHBG ligands from sets of existing commercial substances.

The Glide docking poses of hexestrol, OC and DBP in the zfSHBG ligand-binding site are shown in Figure 3.8. Hexestrol (on the top left, CAS# 84-16-2) is mainly hydrophobic and appears to form hydrogen bonds with residues Lys76, Asn59, and Ser36. OC (top right, CAS# 1139-46-4) appears to hydrogen bond with Asp59 and/or Lys76 while the rest of the OC molecule is hydrophobic; therefore we assume hydrophobic contacts in these areas stabilize the interaction with zfSHBG. The highly aromatic DBP (bottom left, CAS# 3331-46-2) is also hydrophobic and may hydrogen bond with Ser36. Also shown in Figure 3.8 is the Glide docking pose of bisphenol A,

which was identified as a micromolar ligand of zfSHBG (Figure 3.5). Interestingly, the Glide docking poses for hexestrol and bisphenol A (Figure 3.8) indicate that Lys76 provides the main hydrogen bond interaction with hexestrol, while Asp59 that is more likely to hydrogen bond with bisphenol A: an interesting contradictory prediction for two such similar compounds. In a binding assay performed on the K76N zfSHBG mutant, hexestrol almost completely lost its binding affinity while bisphenol A bound to the K76N mutant at even lower concentrations, indicating that the model predictions are correct, with Lys76 representing an important residue for hexestrol binding, while Asp59 is a key component of bisphenol A binding to zfSHBG.

3.5 Discussion

In silico methods identified 25 zfSHBG ligands from the ZINC database, and these could be considered potentially hazardous to aquatic species. In addition, 582 other potential zfSHBG binders identified in silico from the ZINC database likely include bona fide zfSHBG ligands, and these should be studied further. We have also identified 3 non-steroidal ligands of zfSHBG with micromolar affinities out of databases of substances that have been identified as posing environmental concerns, i.e., hexestrol, OC and DBP. Hexestrol is a stilbene estrogen used to promote growth in farm animals that is now banned in many countries because it is a known endocrine disruptor and carcinogen (vom Staal et al., 2005). Interestingly, hexestrol is a derivative of bisphenol A, another stilbene estrogen that is widely used in the plastics industry and a well known endocrine disruptor (Dodds and Lawson, 1938) and carcinogen (Keri et al., 2007), that we have also identified as a micromolar ligand for zfSHBG. OC is a metabolite of an environmentally abundant compound, 4-tert-octylphenol (OP), which is mass produced for use in ink, paints, and varnishes despite the fact that it has

estrogenic activity in fish (Routledge and Sumpter, 1997). Although OC is not produced in industrial quantities, it has been detected in fish (Ferreira-Leach and Hill, 2001), likely because UV irradiation can transform OP into OC (Mazellier and Leverd, 2003). It has been reported that OP also binds to hSHBG (Hodgert-Jury et al., 2000) and its affinity for fish SHBGs should therefore be investigated. DBP is a polycyclic aromatic hydrocarbon (PAH) that is manufactured in relatively low quantities and is used as a reagent in the chemical industry. Like many other PAHs, DBP is not deliberately mass-produced but is a by-product of combustion reactions, and is a derivative of benzo(a)pyrene (Dong et al., 2000), one of the most abundant PAHs. However, benzo(a)pyrene lacks an electronegative oxygen that could participate in hydrogen bonding and likely does not bind SHBG and does not warrant further investigation.

Although we only tested 6 of the top 14 hits out of the 80 000 industrial substances screened for zfSHBG binding in silico, three of them bound in the micromolar range. It is likely therefore that more extensive testing of promising in silico hits would identify other anthropogenic compounds that could bind fish SHBGs with even higher affinity. In this context, it is important to note that recent studies have demonstrated that SHBG in fish gills represents a portal for the uptake of xenobiotics from the aquatic environment (Miguel-Queralt and Hammond, submitted), and that the steroid binding properties of SHBG in different fish can vary considerably because of the relatively poor conservation of SHBG primary structure between species. Thus, although we have validated our model using zfSHBG, it will be important to adapt these methodologies for SHBG in commercially important fish, such as the European sea bass (Miguel-Queralt et al., 2007) and salmonids (Bobe et al., 2008), in which SHBG sequences are known. Nevertheless, the success of this study illustrates the value of

conducting computational toxicology studies on key proteins in multiple species. Virtual screening can be performed rapidly and at low cost, and researchers can now screen almost 70 000 freely available EINECS structures (http://ecb.jrc.it/qsar/information-sources/) for binding to their proteins of interest.

Table 3.1

Experimental pK_a of 5α-dihydrotestosterone (DHT), testosterone, estradiol, androstenedione and ethinylestradiol to hSHBG and zfSHBG.

Molecule	pK_a hSHBG	pK_a zfSHBG
DHT	-9.74	-9.32
Testosterone	-9.20	-9.74
Estradiol	-8.83	-9.30
Androstenedione	-7.46	-9.77
Ethinylestradiol	-6.81	-9.74

Table 3.2

Reproducing previous experimental binding results of estradiol in several hSHBG mutants with ΔG_{bind} from MD Binding free energy calculations.

Mutant	ΔG_{bind} (KJ/mol)	$[^3H]E_2$ bound (cpm)
Wt	-25.83	1800
S42A	-22.66	1000
D65A	-22.68	1300
G58A	-23.23	1100
W84A	-25.3	2500

The '$[^3H]E_2$ bound' column is the binding of $[^3H]$Estradiol to each of the mutants.

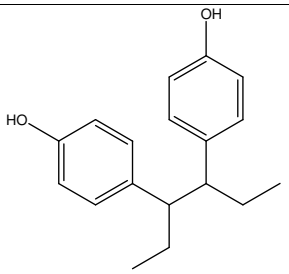
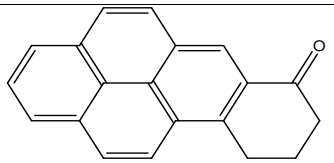
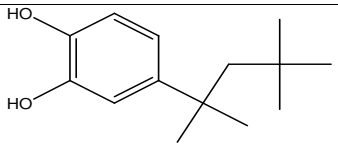
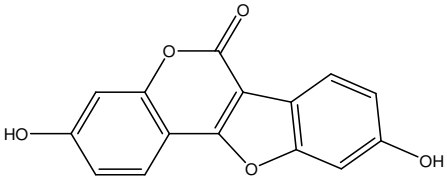
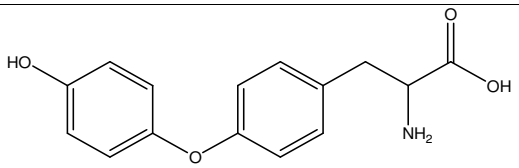
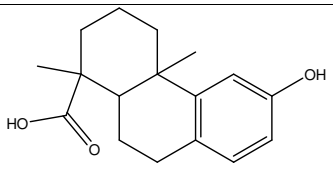
Table 3.3

ΔG_{bind} (KJ/mol) from MD Binding free energy calculations of ethinylestradiol bound to hSHBG mutants.

Mutant	Ethinylestradiol
Wt	-24.6
N82K	-35.42
K134Q	-23.76
M139I	-30.07
V105L	-30.02
M107S	-29.36
L131A	-33.67

Table 3.4

The six commercial substances tested for zfSHBG binding.

CAS	Name	Molecule
84-16-2	Hexestrol*	
3331-46-2	9,10-Dihydrobenzo(a)pyren-7(8H)-one (DBP)*	
1139-46-4	4-tert-octylcatechol (OC)*	
479-13-0	Coumestrol	
1034-10-2	DL-thyronine	
5947-49-9	Podocarpic Acid	

**These three compounds are micromolar zfSHBG ligands.*

Figure 3.1

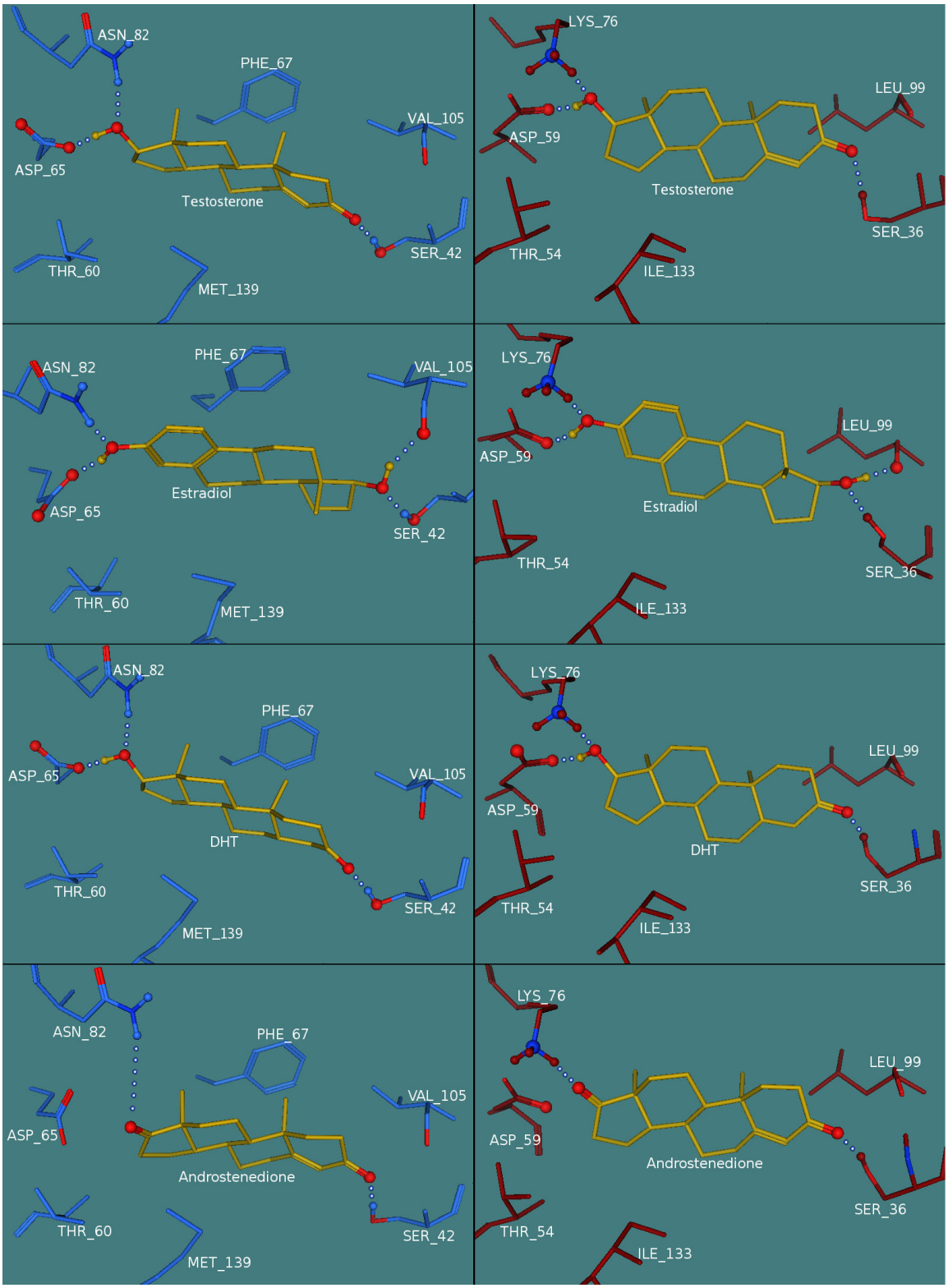
The sequence alignment of hSHBG and zfSHBG used for homology modeling.

```
                                                      +
hSHBG     8  ----DPPAVHLSNGPGQEPIAVMT-FDLTKITKTSSSFEVRTWDPEGVIF  56
zfSHBG    1  DQISGRGTINLAHRQQKWTPAMQTCANLSDIRSIRSFFEFRTLDPEGAVF  50

              + +    + +              +                    +
hSHBG    57  YGDTNPKDDWFMLGLRDGRPEIQLHNHWAQLTVGAGPRLDDGRWHQVEVK  106
zfSHBG   51  YGDTKEGQDWFVLSLRDGIPEMQIGKADILVSVKGGRKLNDGAWHLLELR  100

                                                 +
hSHBG   107  MEGDSVLLEVDGEEVLRLRQVSGPLTSKRHPIMRIALGGLLFPASNLRLP  156
zfSHBG  101  SEGKFVVLEVNNEVELVVGLHSKLAEEQLTGKIRLALGGMLVDKQKLFHP  150

hSHBG   157  LVPALDGCLRRDSWLDKQAEISASAPTSLRSC  188
zfSHBG  151  FEPEMDACIRGGHWLNLSTPWDTDSTWEPRPC  182
```

Only the residues present in the 1KDM template structure and their corresponding zfSHBG amino acids are included. The amino acids of the SHBG steroid binding site within a 3 Å proximity of DHT from 1KDM are marked by a '+'.

Figure 3.2

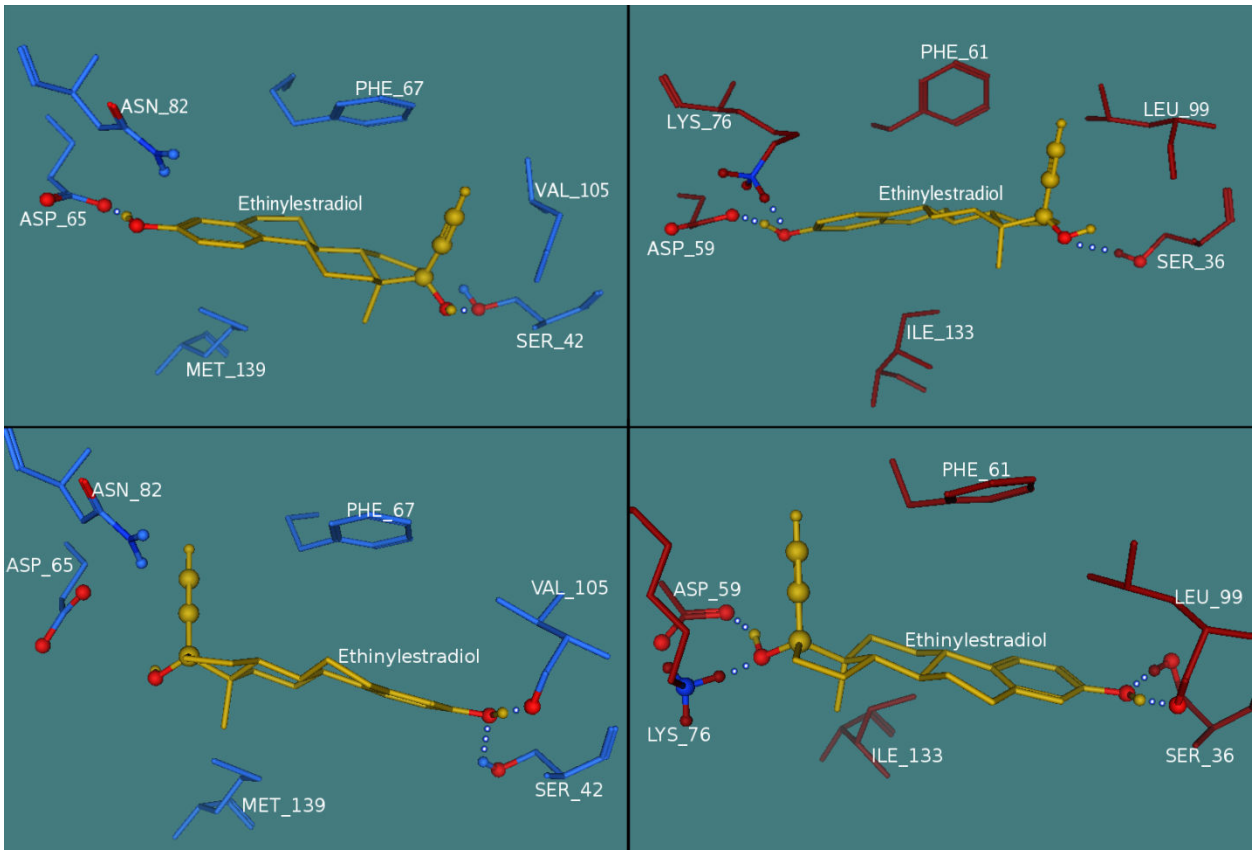
Testosterone, estradiol, DHT, and androstenedione bound to hSHBG and zfSHBG



Only the more important residues for binding are shown. Hydrogen bonds between the ligand and residues within the hSHBG (left, blue) and zfSHBG (right, red) steroid-binding sites are denoted by white dots. Note that the amino acid numbering in hSHBG is shifted six residues from that of zfSHBG so, for instance, Val105 in hSHBG corresponds to Leu99 in zfSHBG.

Figure 3.3

Ethinylestradiol bound to hSHBG and zfSHBG



The top two images show ethinylestradiol bound to hSHBG (left, blue) and zfSHBG (right, red) in the same orientation as estradiol in hSHBG crystal structures. The bottom two images show ethinylestradiol in the pose with the second best binding free energy estimate. Note that the amino acid numbering in hSHBG is shifted six residues from that of zfSHBG so, for instance, Val105 in hSHBG corresponds to Leu99 in zfSHBG.

Figure 3.4

Scatter plots of GS versus pK_a for hSHBG and GS versus pK_a for zfSHBG

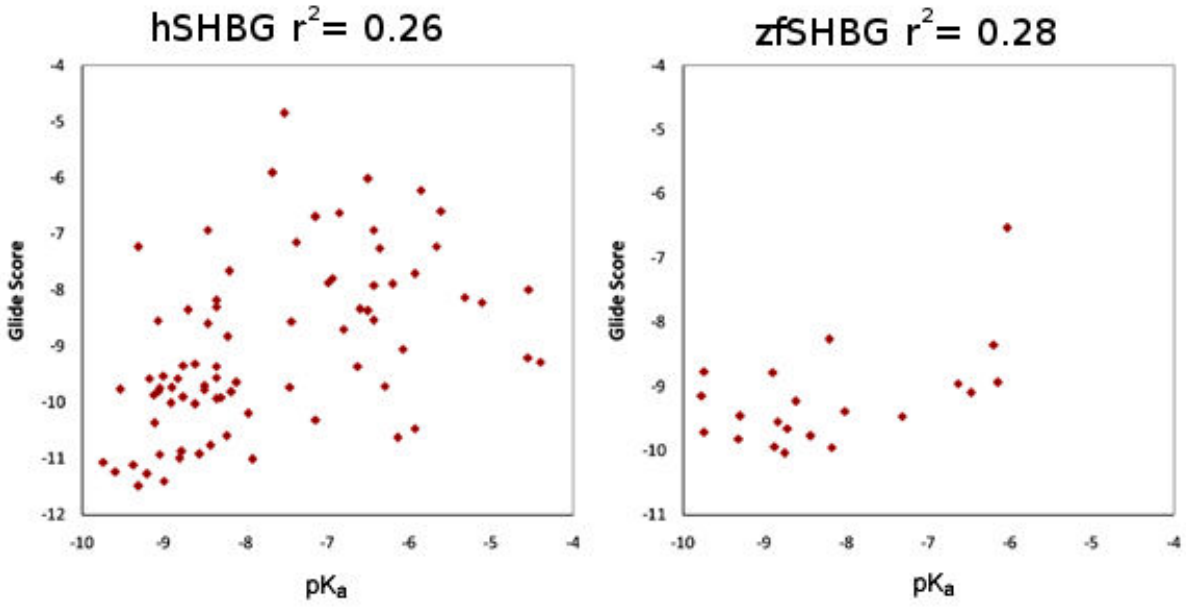
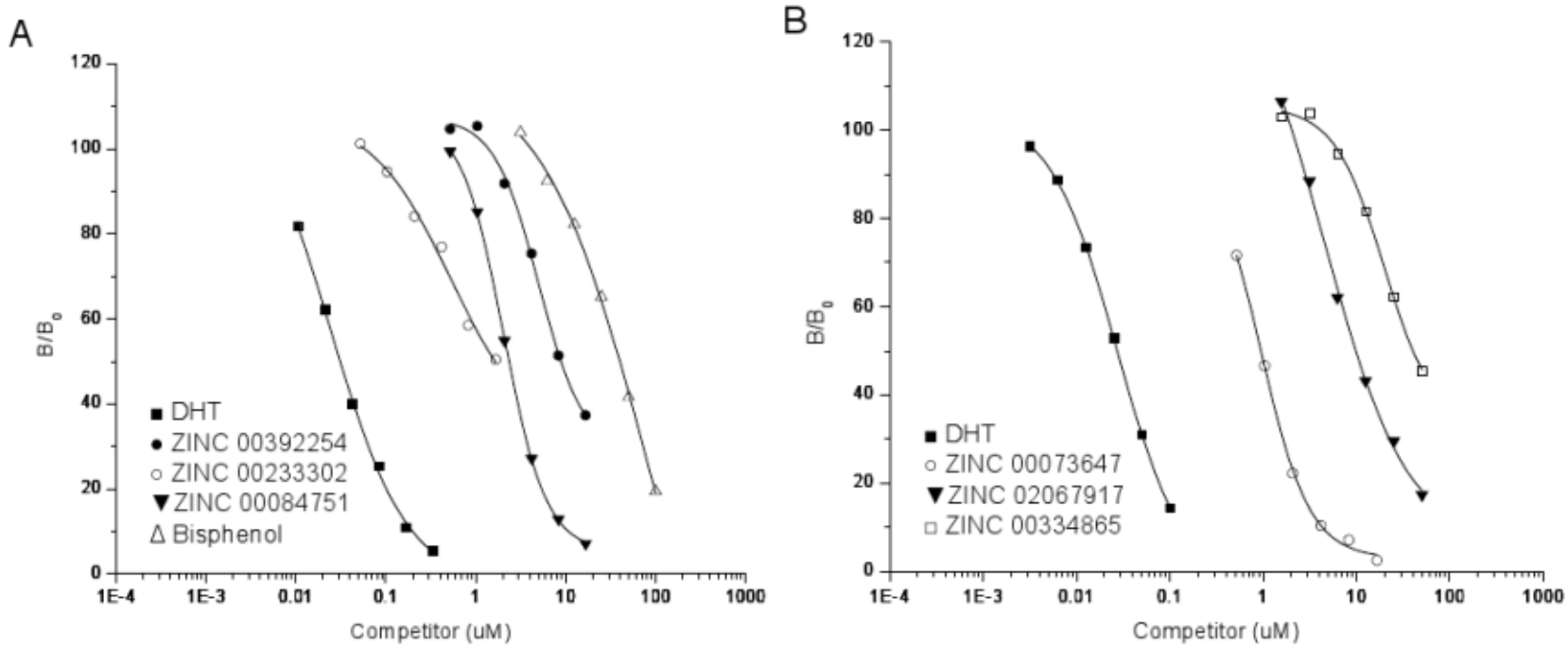


Figure 3.5

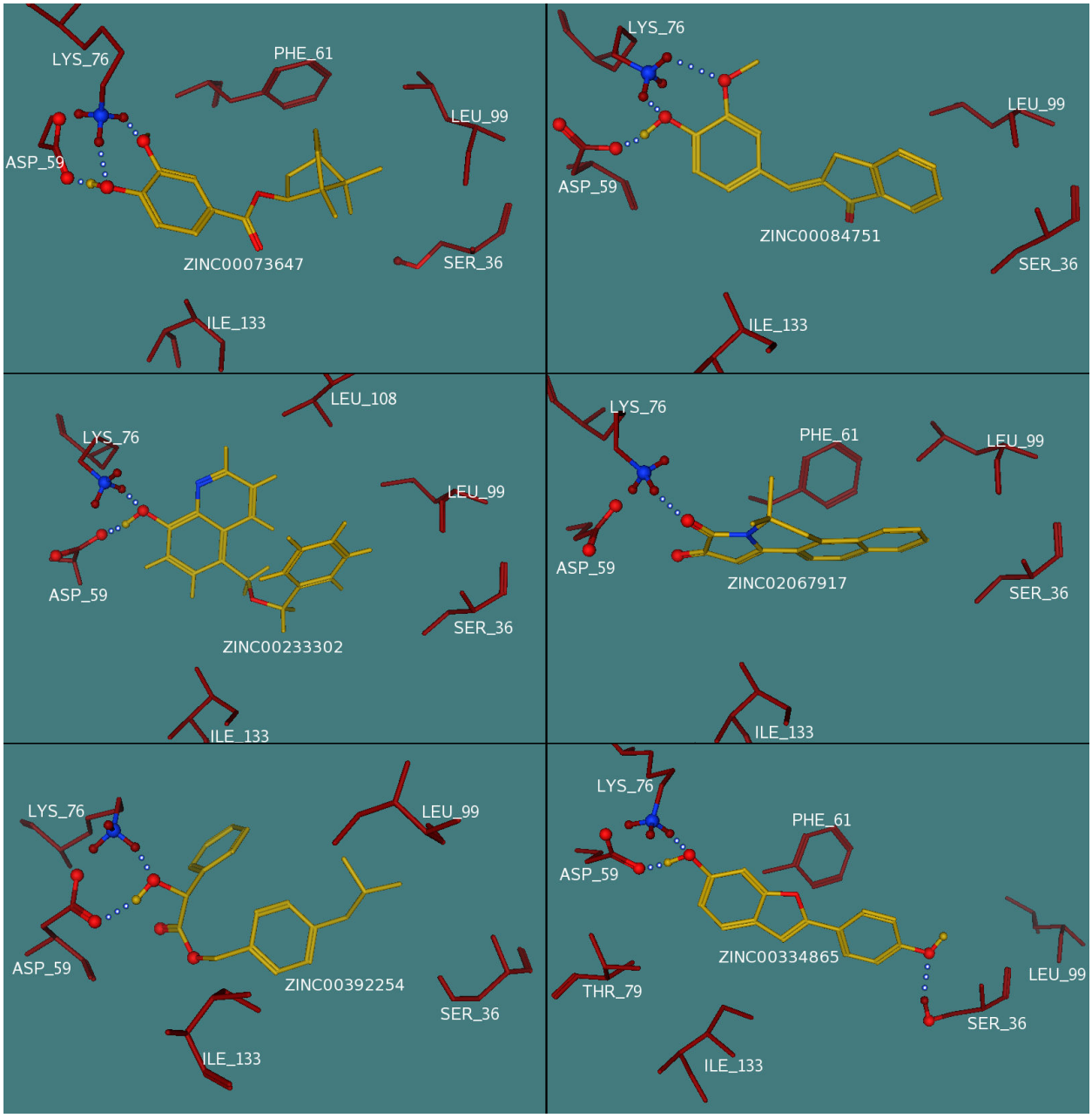
The displacement curves for test compounds used in the in vitro competition assay to determine the relative binding affinities of zfSHBG ligands.



Amounts of [^{3}H] DHT bound to zfSHBG in the presence of increasing concentrations of competitors.

Figure 3.6

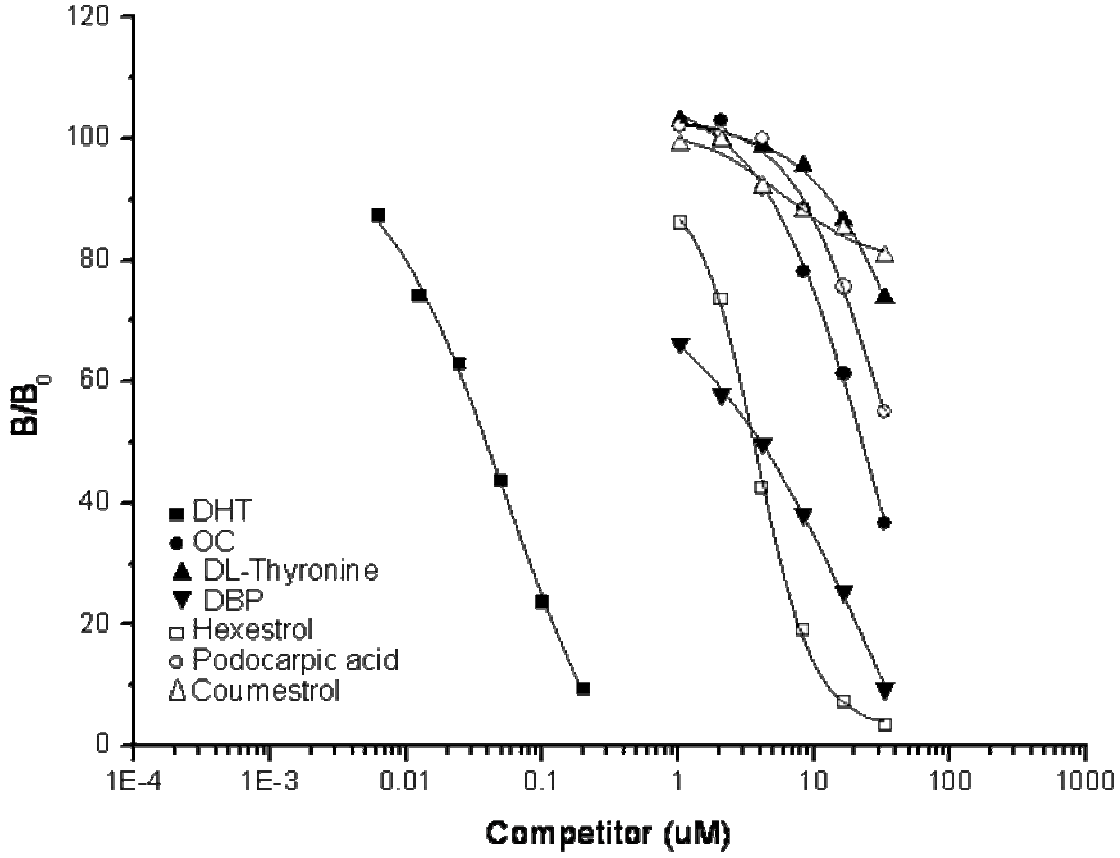
Six ZINC compounds docked in the zfSHBG binding pocket.



Only the residues which are most relevant to ligand binding and which do not interfere with the view of the ligand are shown. Hydrogen bonds between the ligand and residues within the zfSHBG steroid-binding site are represented as white dots.

Figure 3.7

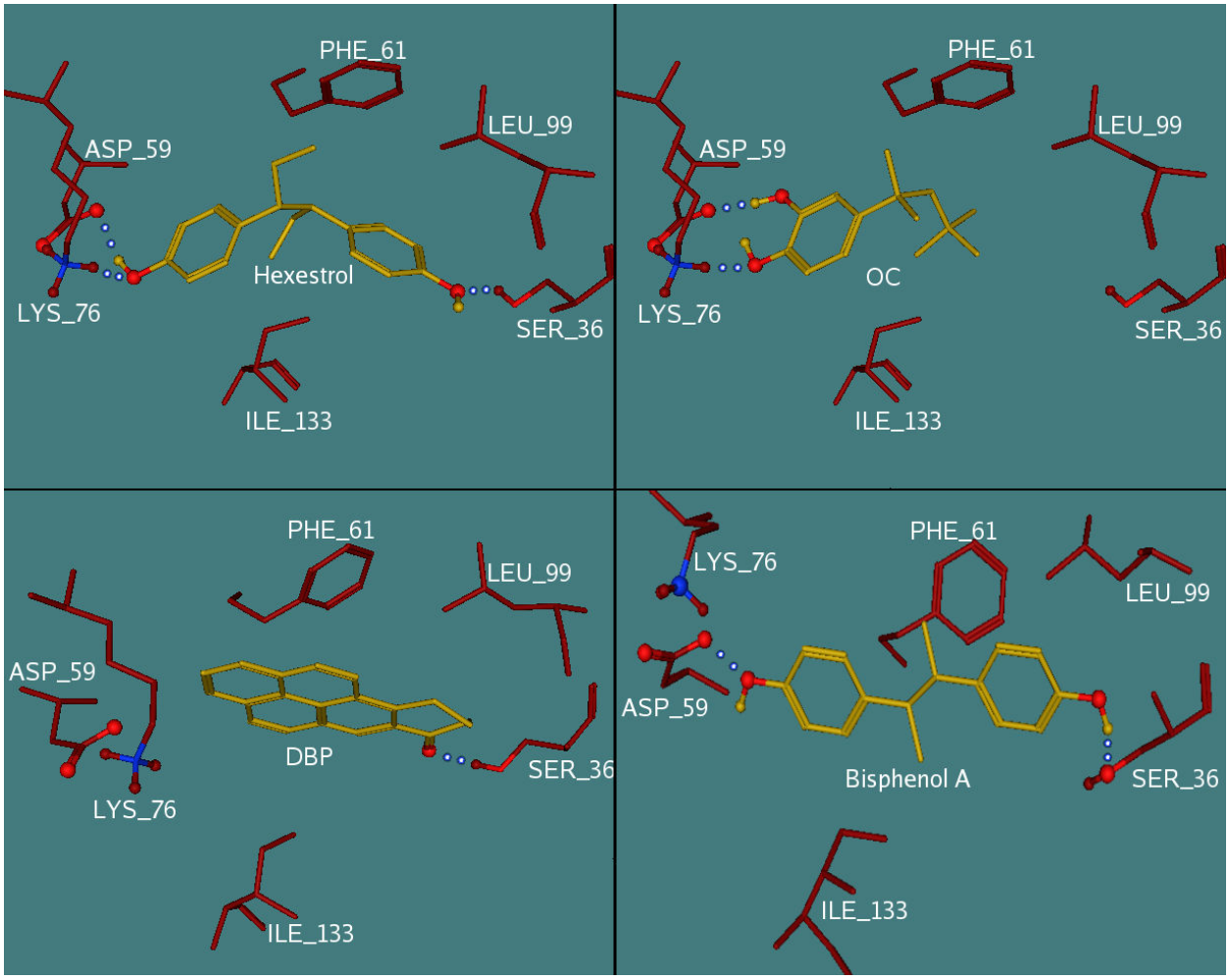
The displacement curves of the four binders used in the competition assay to determine the relative binding affinities of zfSHBG ligands.



Amounts of [^{3}H] DHT bound to zfSHBG in the presence of increasing concentrations of competitors.

Figure 3.8

Hexestrol, OC, DBP and Bisphenol A docked in the zfSHBG binding pocket.



Only the residues which are most relevant to ligand binding and which do not interfere with the view of the ligand are shown. Hydrogen bonds between the ligand and residues within the zfSHBG steroid-binding site are represented by the blue and white dots.

3.6 References

Asikainen, A. H., Ruuskanen, J., and Tuppurainen, K. A. (2004). Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds. *SAR QSAR Environ Res.* **15**, 19-32.

Blomquist, C. H., Lima, P. H., and Hotchkiss, J. R. (2005). Inhibition of 3alpha-hydroxysteroid dehydrogenase (3alpha-HSD) activity of human lung microsomes by genistein, daidzein, coumestrol and C(18)-, C(19)- and C(21)-hydroxysteroids and ketosteroids. *Steroids* **70**, 507-514.

Bobe, J., Mahe, S., Nguyen, T., Rime, H., Vizziano, D., Fostier, A., and Guiguen, Y. (2008). A novel, functional and highly divergent sex hormone-binding globulin that may participate in the local control of ovarian functions in salmonids. *Endocrinology*. Mar 13 [Epub ahead of print]

Chemical Computing Group, Inc. (2006). Molecular Operating Environment (MOE), Version 2006.08**.** Montreal, Canada

Chemical Computing Group, Inc. (2005). SVL exchange: http://svl.chemcomp.com/viewcat.php

Cherkasov, A., Ban, F., Santos-Filho, O., Thorsteinson, N., Fallahi, M., and Hammond, G. L. (2008). An updated steroid benchmark set and its application in the discovery of novel nanomolar ligands of sex hormone-binding globulin. *J Med Chem* **51**, 2047-2056.

Cherkasov, A., Shi, Z., Fallahi, M., and Hammond, G. L. (2005). Successful in silico discovery of novel nonsteroidal ligands for human sex hormone binding globulin. *J. Med. Chem.* **48**, 3203-3213.

Cramer, R. D., Patterson, D. E., and Bunce, J. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110**, 5959-5967.

Dodds, E.C., Lawson, W., Noble, R.L. (1938). Biological effects of the synthetic oestrogenic substance 4: 4'-dihydroxy-a: B-dimethylstilbene. *Lancet.* **235**, 1389-1391

Dong, S., Hwang, H. M., Harrison, C., Holloway, L., Shi, X., and Yu, H. (2000). UVA light-induced DNA cleavage by selected polycyclic aromatic hydrocarbons. *Bull. Environ. Contam. Toxicol.* **64**, 467-474.

Environment Canada. (2006). The Canadian environmental protection act environmental registry. http://www.ec.gc.ca/CEPARegistry/default.cfm

Erbs, M., Hoerger, C. C., Hartmann, N., and Bucheli, T. D. (2007). Quantification of six phytoestrogens at the nanogram per liter level in aqueous environmental samples using 13C3-labeled internal standards. *J. Agric. Food Chem.* **55**, 8339-8345.

European Chemicals Bureau. (2002). The European inventory of existing commercial substances (EINECS). http://ecb.jrc.it/existing-chemicals/

Ferreira-Leach, A. M., and Hill, E. M. (2001). Bioconcentration and distribution of 4-tert-octylphenol residues in tissues of the rainbow trout (Oncorhynchus mykiss). *Mar. Environ. Res.* **51**, 75-89.

Gasteiger J., Marsili M. (1980). Iterative partial equalization of orbital electronegativity – a rapid access to atomic charges. *Tetrahedron.* **26** 3219.

Grishkovskaya, I., Avvakumov, G. V., Hammond, G. L., Catalano, M. G., and Muller, Y. A. (2002a). Steroid ligands bind human sex hormone-binding globulin in specific orientations and produce distinct changes in protein conformation. *J. Biol. Chem.* **277**, 32086-32093.

Grishkovskaya, I., Avvakumov, G. V., Hammond, G. L., and Muller, Y. A. (2002b). Resolution of a disordered region at the entrance of the human sex hormone-binding globulin steroid-binding site. *J. Mol. Biol.* **318**, 621-626.

Halgren, T.A. (1996). The Merck Force Field. *J. Comp. Chem.* **17** 490-512, 520-552, 553-586, 587-615, 616-641.

Hammond, G. L., Avvakumov, G. V., and Muller, Y. A. (2003). Structure/function analyses of human sex hormone-binding globulin: effects of zinc on steroid-binding specificity. *J. Steroid Biochem. Mol. Biol.* **85**, 195-200.

Hansson, T., Marelius, J., and Aqvist, J. (1998). Ligand binding affinity prediction by linear interaction energy methods. *J. Comput. Aided Mol. Des.* **12**, 27-35.

Hodgert Jury, H., Zacharewski, T. R., and Hammond, G. L. (2000). Interactions between human plasma sex hormone-binding globulin and xenobiotic ligands. *J. Steroid Biochem. Mol. Biol.* **75**, 167-176.

Irwin, J. J., and Shoichet, B. K. (2005). ZINC -a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model* **45**, 177-182.

Keri, R. A., Ho, S. M., Hunt, P. A., Knudsen, K. E., Soto, A. M., and Prins, G. S. (2007). An evaluation of evidence for the carcinogenic activity of bisphenol A. *Reprod. Toxicol.* **24**, 240-252.

Klebe, G., Abraham, U., and Mietzner, T. (1994). Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **37**, 4130-4146.

Korhonen, S. P., Tuppurainen, K., Laatikainen, R., and Perakyla, M. (2003). FLUFF-BALL, a template-based grid-independent superposition and QSAR technique:

validation using a benchmark steroid data set. *J. Chem. Inf. Comput. Sci.* **43**, 1780-1793.

Liu, S. S., Yin, C. S., Li, Z. L., and Cai, S. X. (2001). QSAR study of steroid benchmark and dipeptides based on MEDV-13. *J. Chem. Inf. Comput. Sci.* **41**, 321-329.

Liu, S. S., Yin, C. S., and Wang, L. S. (2002). Combined MEDV-GA-MLR method for QSAR of three panels of steroids, dipeptides, and COX-2 inhibitors. *J. Chem. Inf. Comput. Sci.* **42**, 749-756.

Matthews, J., Celius, T., Halgren, R., and Zacharewski, T. (2000). Differential estrogen receptor binding of estrogenic substances: a species comparison. *J. Steroid Biochem. Mol. Biol.* **74**, 223-234.

Mazellier, P., and Leverd, J. (2003). Transformation of 4-tert-octylphenol by UV irradiation and by an H2O2/UV process in aqueous solution. *Photochem. Photobiol. Sci.* **2**, 946-953.

Miguel-Queralt, S., Blazquez, M., Piferrer, F., and Hammond, G. L. (2007). Sex hormone-binding globulin expression in sea bass (Dicentrarchus labrax L.) throughout development and the reproductive season. *Mol. Cell Endocrinol.* **276**, 55-62.

Miguel-Queralt, S., Knowlton, M., Avvakumov, G. V., Al-Nouno, R., Kelly, G. M., and Hammond, G. L. (2004). Molecular and functional characterization of sex hormone binding globulin in zebrafish. *Endocrinology* **145**, 5221-5230.

Petra, P. H., Namkung, P. C., Senear, D. F., McCrae, D. A., Rousslang, K. W., Teller, D. C., and Ross, J. B. (1986). Molecular characterization of the sex steroid binding protein (SBP) of plasma. Re-examination of rabbit SBP and comparison with the human, macaque and baboon proteins. *J. Steroid Biochem.* **25**, 191-200.

Routledge, E. J., and Sumpter, J. P. (1997). Structural features of alkylphenolic chemicals associated with estrogenic activity. *J. Biol. Chem.* **272**, 3280-3288.

Schuttelkopf, A. W., and van Aalten, D. M. (2004). PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta. Crystallogr. D. Biol. Crystallogr.* **60**, 1355-1363.

Schrödinger, Inc. (2006). Glide, version 4.0. San Diego, CA.

Schrödinger, Inc. (2004). Maestro. San Diego, CA.

Scott A. P., Pinillos M. L., Huertas M. (2005). The rate of uptake of sex steroids from water by *Tinca tinca* is influenced by their affinity for sex steroid binding protein plasma. *J. Fish. Biol.* **67** 182-200.

Siiteri, P. K., Murai, J. T., Hammond, G. L., Nisker, J. A., Raymoure, W. J., and Kuhn, R. W. (1982). The serum transport of steroid hormones. *Recent Prog. Horm. Res.* **38**, 457-510.

Tripos, Inc. (2006). SYBYL, version 7.2. St. Louis, MO.

Tuppurainen, K., Viisas, M., Laatikainen, R., and Perakyla, M. (2002). Evaluation of a novel electronic eigenvalue (EEVA) molecular descriptor for QSAR/QSPR studies: validation using a benchmark steroid data set. *J. Chem. Inf. Comput. Sci.* **42**, 607-613.

Tuppurainen, K., Viisas, M., Perakyla, M., and Laatikainen, R. (2004). Ligand intramolecular motions in ligand-protein interaction: ALPHA, a novel dynamic descriptor and a QSAR study with extended steroid benchmark dataset. *J. Comput. Aided Mol. Des.* **18**, 175-187.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. (2005). GROMACS: fast, flexible, and free. *J. Comput. Chem.* **26**, 1701-1718.

Vom Saal, F. S., Richter, C. A., Ruhlen, R. R., Nagel, S. C., Timms, B. G., and Welshons, W. V. (2005). The importance of appropriate controls, animal feed, and animal models in interpreting results from low-dose studies of bisphenol A. *Birth Defects Res. A. Clin. Mol. Teratol.* **73**, 140-145.

4 CONCLUDING REMARKS

4.1 Goals accomplished

This work presents an example of modern computer-aided technology encompassing theoretical and experimental method development and drug discovery components.

As a result of this work, we have updated the popular steroid benchmark set representing one of the most broadly used ‘golden standard’ data set for computational drug discovery. The set has been expanded with a number of steroidal and non-steroidal hSHBG ligands, moreover, the steroid binding orientations have been modified to better reflect recent crystallographic observations (Grishkovskaya et al., 2002).

The updated and corrected benchmark set alignments were used to test the sensitivity of the CoMFA (Cramer et al., 1988) and CoMSIA (Klebe et al., 1994) methods to ligand orientation and we found that models constructed with the updated alignments were almost indistinguishable from previous models and, therefore, we concluded that the CoMFA and CoMSIA methods can overlook differences in steroid ligand alignment.

In order to identify hSHBG ligands, we developed a QSAR model consisting of inductive QSAR descriptors (Cherkasov et al., 2005; Cherkasov, 2005) using the updated benchmark training set. This model, combined with the corrected CoMFA and CoMSIA solutions, and molecular docking were applied in a virtual screen of the large ZINC chemical database (Irwin and Shoichet, 2005) for hSHBG ligands. Overall, eight novel non-steroidal high-affinity hSHBG ligands were identified.

Furthermore, to gain more insight into the mechanism of the receptor-ligand interactions, we computed binding free energy simulations on hSHBG mutant models

which predicted that ethinylestradiol has a higher affinity for the hSHBG N82K mutant than for wildtype hSHBG. From this, we deduced that the corresponding zfSHBG Lys76 residue contributes to the strong binding of ethinylestradiol to zfSHBG. A binding assay involving the zfSHBG K76N mutant demonstrated that ethinylestradiol has a lower affinity for this mutant than for wildtype zfSHBG. This confirmed the binding free energy prediction that zfSHBG Lys76 contributes to the strong binding of ethinylestradiol to zfSHBG.

This knowledge helped in creating a powerful multi-method discovery pipeline combining docking, CoMFA, and CoMSIA predictions which, when applied to a zfSHBG homology model, identified six zfSHBG ligands from an existing commercial substances database, three of which bind to zfSHBG in the micromolar range, and are therefore potential threats to aquatic species. In silico methods identified 25 zfSHBG ligands from the ZINC database, and these could be considered potentially hazardous to aquatic species. In addition, 582 other potential zfSHBG binders identified in silico from the ZINC database likely include bona fide zfSHBG ligands.

4.2 Future directions

The high-affinity non-steroidal hSHBG ligands identified are leads for potential therapeutic agents, therefore, lead optimization studies of these compounds should ensue to design modified substances with even higher affinity for hSHBG. For instance, since six of the eight high-affinity non-steroidal hSHBG ligands identified lack an electronegative atom to participate in hydrogen bonding with Ser42 of hSHBG, adding substituents such as a hydroxyl group to this area of these molecules may increase their affinity for hSHBG and may cause them to be more hSHBG-specific, reducing the likelihood of causing side-effects.

The in silico screen of existing commercial chemicals identified hundreds of potential zfSHBG ligands, but only six were tested for zfSHBG binding, therefore, more of these hits should be tested in order to possibly expose some high volume chemicals, or their derivatives, as zfSHBG ligands. Any experimentally validated high-affinity zfSHBG ligands from the existing commercial substances should be studied further in more thorough aquatic toxicity tests. Also, as mentioned in chapter 3, section 5, the anthropogenic substance database should be screened for ligands of commercially important fish such as the European sea bass (Miguel-Queralt et al., 2007) and salmonids (Bobe et al., 2008), for which SHBG sequences are known.

Virtual screening of existing chemicals should be applied to other biologically relevant proteins in order to potentially discover toxic effects of commercial substances. The publishing of a version of the manuscript in chapter three will encourage researchers to perform similar computational toxicology studies on their proteins of interest.

4.3 References

Bobe, J., Mahe, S., Nguyen, T., Rime, H., Vizziano, D., Fostier, A., and Guiguen, Y. (2008). A novel, functional and highly divergent sex hormone-binding globulin that may participate in the local control of ovarian functions in salmonids. *Endocrinology*. Mar 13 [Epub ahead of print]

Cherkasov, A., Shi, Z., Li, Y., Jones, S. J., Fallahi, M., and Hammond, G. L. (2005). 'Inductive' charges on atoms in proteins: comparative docking with the extended steroid benchmark set and discovery of a novel SHBG ligand. *J. Chem. Inf. Model.* **45**, 1842-1853.

Cherkasov A. (2005). 'Inductive' Descriptors. 10 Successful Years in QSAR. *Curt. Comp-Aided Drug Design.* **1**, 21-42.

Cramer, R. D., Patterson, D. E., and Bunce, J. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110**, 5959-5967.

Grishkovskaya, I., Avvakumov, G. V., Hammond, G. L., Catalano, M. G., and Muller, Y. A. (2002). Steroid ligands bind human sex hormone-binding globulin in specific orientations and produce distinct changes in protein conformation. *J. Biol. Chem.* **277**, 32086-32093.

Irwin, J. J., and Shoichet, B. K. (2005). ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177-182.

Klebe, G., Abraham, U., and Mietzner, T. (1994). Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **37**, 4130-4146.

Miguel-Queralt, S., Blazquez, M., Piferrer, F., and Hammond, G. L. (2007). Sex hormone-binding globulin expression in sea bass (Dicentrarchus labrax L.) throughout development and the reproductive season. *Mol. Cell. Endocrinol.* **276**, 55-62.