January 2015

Development and Application of Pseudoreceptor Modeling

Gregory Lee Wilson
Purdue University

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Gregory L. Wilson

Entitled
Development and Application of Pseudoreceptor Modeling

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Markus Lill Chair	Jean-Christophe Rochet
Chiwook Park	Daisuke Kihara

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Markus Lill

Approved by: Jean-Christophe Rochet 11/24/2015
Head of the Departmental Graduate Program Date

DEVELOPMENT AND APPLICATION OF PSEUDORECEPTOR METHODS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Gregory L. Wilson

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2015

Purdue University

West Lafayette, Indiana

To my parents who have supported me through this process

Gerald Wilson and Constance Wilson

ACKNOWLEDGEMENTS

I first wish to acknowledge Dr. Markus Lill. I would not be where I am without his support and guidance not just professionally, but in all aspects of being a researcher and his support of me in my personal struggles. He has been an excellent mentor, knowing when to guide me, and when to push me, and I truly appreciate his efforts.

Next, I wish to thank all my fellow Lill Group members, past and present. They were always willing to answer a question, lend a helping hand, or work together to find solutions to our problems, common or not. I specifically wits to thank Dr. Laura Kingsley and Morgan Essex for their efforts in our collaboration on Cytochrome P450.

I also wish to thank my other collaboration partners, namely the late and missed Dr. Gibbs and members of his lab. The opportunity to provide computational support in a true drug design scenario and earn a patent as part of my education was a great experience that I feel will serve me well in the future.

Finally, I wish to acknowledge the efforts of my committee: Dr. Rochet, Dr. Kihara, and Dr. Park. Their questions and challenges during our interactions helped teach me to think and act as a scientist and a researcher, to consider what I do and do not know, and what I should know. They also provided valuable insight into effectively communicating my research.

TABLE OF CONTENTS

Page

Page

Page

Page

LIST OF TABLES

Table Page

Appendix Table

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

ABSTRACT

Wilson, Gregory L. Ph.D., Purdue University, December 2015. Development And Application Of Pseudoreceptor Methods. Major Professor: Markus Lill.

Quantitative Structure-Activity Relationship (QSAR) methods are a commonly used tool in the drug discovery process. These methods attempt to form a statistical model that relates descriptor properties of a ligand to the activity of that ligand compound towards a specific desired physiological response. QSAR methods are known as a ligand-based method, as they specifically use information from ligands and not protein structural data. However, a derivation of QSAR methods are pseudoreceptor methods. Pseudoreceptor methods go beyond standard QSAR by building a model representation of the protein pocket. However, the ability of pseudoreceptors to accurately replicate natural protein surfaces has not been studied. The goal of this thesis work is to investigate the necessary descriptors to map a protein binding pocket and a method to accurately recreate the 3-D spatial structure of the binding pocket. In addition, additional applications of existing pseudoreceptor methods are explored.

To identify the necessary descriptors to map a protein binding pocket, we developed a program that decomposes the protein-ligand interaction surface from a large number of ligand-bound protein crystal structures. The binding pockets of the protein structure are identified, and then the physico-chemical properties of the protein are mapped onto the solvent accessible surface of the binding pocket. A number of 2-D Gaussian functions are then placed onto this surface to model the protein's physico-chemical properties. We found that a small number of these Gaussians were able to accurately replicate the properties of the protein.

With this knowledge, we then desired a means of accurately recreating the binding pocket surfaces of proteins only the structures of their bound ligands. Typically in pseudoreceptor methods either the average or combined solvent accessible surface of the ligand set is used. To test this, we generated iso-level surfaces of the solvent accessible surfaces of sets of ligands for which the co-crystallized protein structure is available. We also tested additional sets of surfaces located beyond the ligand's solvent accessible surface. We found that any single surface was unable to accurately reproduce the protein-ligand interaction surface, and multi-surface approach using numerous iso-surfaces is needed to accurately represent the protein.

Finally, we explored the application of RAPTOR, an existing pseudoreceptor method, to the problem of the prediction of Sites-of-Metabolism (SoM) for Cytochrome P450s (Cyps). In our approach, we used RAPTOR as a means of discriminating between active (correctly predicted SoM) docking poses of ligands from decoy (incorrect SoM) poses. With our method, we achieved the highest reported rate of SoM prediction across nine Cyp isoforms, with the best reported performance on seven of those nine isoforms.

CHAPTER 1. INTRODUCTION

1.1 Combined Ligand-based and Structure-based Computer-Aided Drug Design

The majority of Computer-Aided Drug Design (CADD) methods can be divided into two categories: ligand-based drug design, and structure-based drug design. These categories are named after the origin of the data used in the design procedure. Ligand-based drug design efforts are based off the analysis of the biological activities and chemical properties of a set of ligands, and are often used when little to no information about the structure of the target protein is available. A primary example of ligand-based drug design would be the wide variety of Quantitative Structure Activity Relationship (QSAR) techniques. On the other hand, when there is sufficient information about the three-dimensional structure of the target protein, especially if an X-ray structure is available, structure-based drug design methods are routinely applied in the drug development process. These techniques focus on simulating the interactions of potential ligands with the protein structure. Molecular dynamics or Monte Carlo simulation-based free energy methods and protein-ligand docking simulations are major types of structure-based design techniques. However, as the number of available protein-ligand crystal structures continues to rise, and as more and more physicochemical and biological data for ligands is published, there is an increasing number of systems where both ligand and protein structure data is available. Thus, there is a growing trend of attempting to perform both ligand-based and structure-based drug design on the same protein system. These efforts may be as simple as performing QSAR or pharmacophore studies and docking on the same system, and there are a number of examples of such occurrences in the literature[1,2]. What we are focused on here, however, are integrated methods of combining ligand-based and structure-based drug design concepts into a single technique.

Some of the earliest work on combining techniques from structure- and ligand-based design was the adaptation of the GRID program[3,4] to ligand-based design leading to the GRID-GOLPE approach[5]. The GRID method can be used on a protein structure to identify hotspots of possible protein-ligand interactions, e.g. favorable interactions with hydrogen-bonding or hydrophobic groups. In the GRID-GOLPE adaptation, GRID is applied on a set of ligand structures binding to a common binding site. GOLPE[6] performs the chemometric analysis by identifying the descriptors strongly correlating with biological activity and generating a multivariate regression using those descriptors. The methods that we will discuss in this section cover two major categories where significant development of integrated structure-based and ligand-based drug design is occurring: interaction-based methods, like GRID-GOLPE, and docking-similarity based methods (Figure 1.1).

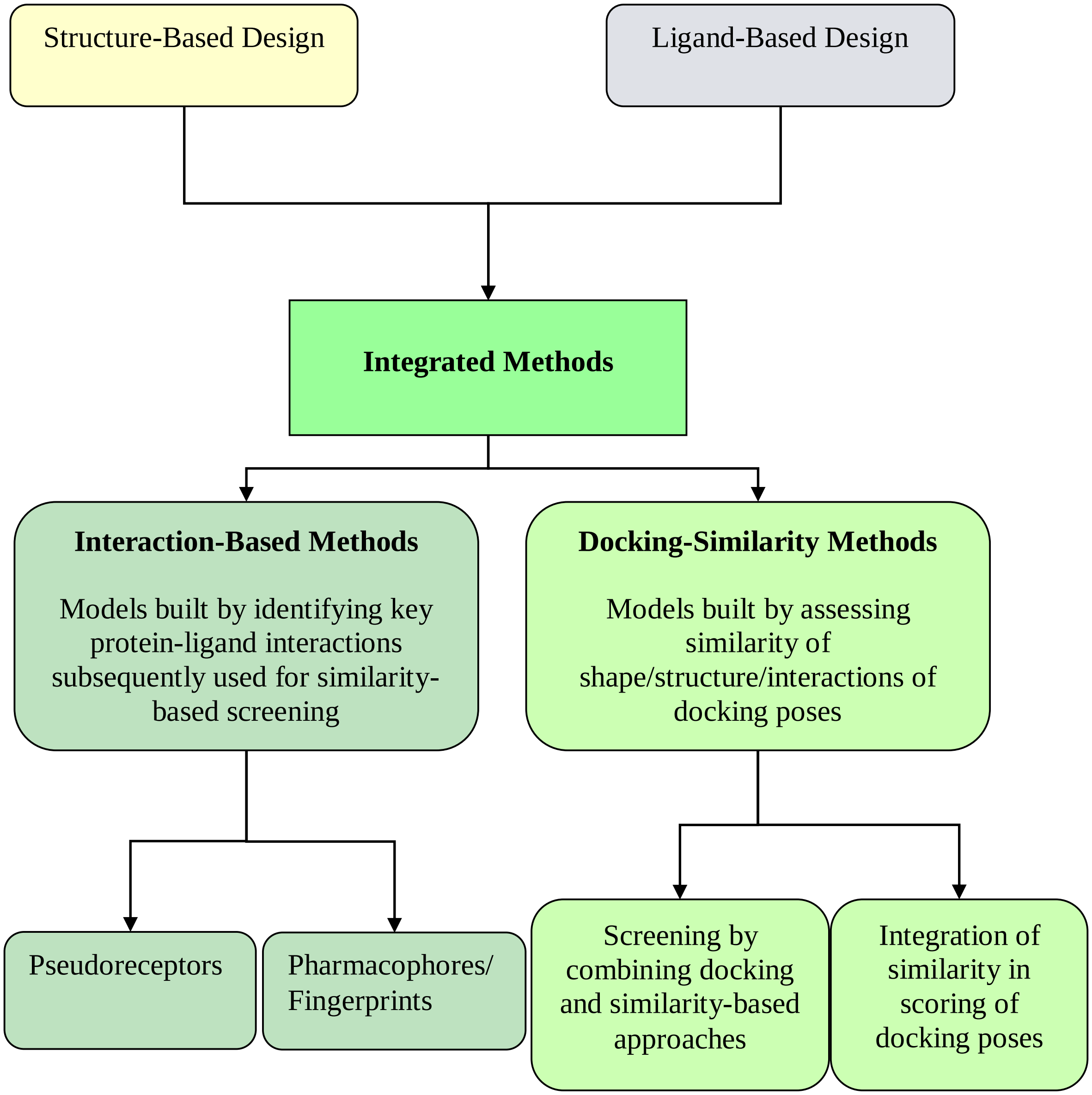

Structure-Based Design
Ligand-Based Design
Integrated Methods
Interaction-Based Methods
Models built by identifying key protein-ligand interactions subsequently used for similarity-based screening
Docking-Similarity Methods
Models built by assessing similarity of shape/structure/interactions of docking poses
Pseudoreceptors
Pharmacophores/ Fingerprints
Screening by combining docking and similarity-based approaches
Integration of similarity in scoring of docking poses

Figure 1.1: Classification scheme of integrated structure and ligand-based methods. The major classification into two major categories includes interaction-based and docking similarity-based methods. Each of those categories contain two subcategories: pseudoreceptor methods and pharmacophore/fingerprint-based methods for the interaction-based methods, and combined structure-ligand based virtual screening approaches and methods that integrate similarity-based concepts into the scoring process of ligand docking.

1.1.1 Interaction-Based Methods

One major class of methods integrating both ligand-based and structure-based drug design methods is based on comparing or modeling protein-ligand interactions across similar protein-ligand systems. These concepts seek to identify key protein-ligand interactions from known data and utilize this interaction data to identify ligands with similar interaction profile. This class of integrated methods can be further divided into two sub-categories (Figure 1.1). The first sub-category is the pseudoreceptor techniques that correlate similarities between ligands with measured biological activity, similar to QSAR, but then use this data to establish a structural representation of the protein-ligand binding pocket[11,45,46,48,50-72]. The other set of techniques is the converse of the first category. These methods analyze protein-ligand interactions in structural data to extract key types of interactions, and then translate that information into a simplified mathematical representation that can be used by similarity-based methods to screen for active lead compounds in ligand libraries[73-96]. Many techniques from this category are based upon fingerprint or pharmacophore models.

1.1.2 Docking and Screening Based Methods

The second major class of integrated structure-based and ligand-based design techniques is those methods which combine structure-based docking techniques with ligand-based similarity information (Figure 1.1)[2,97-104]. The first subcategory is screening-based methods. These methods use ligand similarity to aid high-throughput virtual screening in one of two ways. When there is a known hit or lead compound, similarity studies are used to enrich ligand libraries to reduce the number of compounds that are docked. The other approach is to use docking to identify a possible hit, and then screen a ligand library for similar ligands as alternative hits. The other category addresses one of the major known issues with docking, the scoring problem, by integrating ligand similarity directly into the scoring process.

1.2 Pseudoreceptor Methods

As mentioned previously, pseudoreceptor methods are a means of integrating structure-based and ligand-based techniques. Pseudoreceptor[7] methods are primarily expansions of Quantitative Structure Activity Relationship (QSAR) techniques, mainly 3-D-QSAR techniques such as CoMFA[8], CoMSIA[9,10], or GOLPE[6], that place physicochemical information onto 3-D space surrounding a set of aligned reference compounds that bind into the same binding site of a common macromolecular target. Pseudoreceptor methods expand this mapping by attempting to create models of the target protein binding site around the ligand ensemble. These representative pseudoreceptor models are intended to contain key protein-ligand interactions, and to map these interactions into an appropriate shape and volume.

The aim of generating these models is to be able to rationally modify or propose new small molecules that are complementary to the pseudoreceptor model and to accurately predict binding affinities for a series of potential ligands. Early pseudoreceptor methods involved the manual folding of peptide chains around the ligand ensemble[11], but these methods have now been expanded into a wide-variety of automated computational methods. There are several major classes of pseudoreceptor methods including atom-based, surface-based, fragment-based and residue-based methods[1-6].

1.2.1 Challenges of Pseudoreceptor Methods

Two critical factors in the overall process of pseudoreceptor modeling are the chemical space of the ligand set and the ligand alignment process. The chemical space of a ligand set refers to the set of physicochemical properties present in the entire ligand library and the span of related binding affinities. The pseudoreceptor model can only account for those features present in the chemical space of the ligand library, e.g. if a protein has a hydrogen-bonding residue in the binding pocket with no matching functional group in the ligand set, the pseudoreceptor model will lack that particular hydrogen-bonding feature.

The alignment of the ligand set plays an important role in generating the pseudoreceptor model as well. In order to accurately represent the 3-D structure of the protein-binding pocket, the correct ligand binding mode is necessary. This is a non-trivial challenge, especially with regards to highly flexible ligands. As such, a large number of methods for alignment have been developed and utilized for the various pseudoreceptor methods. Alignment techniques include pharmacophore based methods, molecular simulations, other similarity-based methods, as well as docking methods if protein structure information is available[12-21].

1.2.2 Surface-based methods

One major class of pseudoreceptors is surface-based methods, where the pseudoreceptor is represented as a curved 3-D surface with physicochemical properties mapped onto it representing protein properties important for protein-ligand interactions[44-48]. These surfaces are generated in a number of ways. In Receptor Surface Models (RSM), a “shape field” for each ligand is generated that represents the molecular volume[45-47]. The fields for all ligands are then combined, and an iso-level surface is generated based on the combined shape field. In RAPTOR, an iso-surface approximating the solvent-accessible surface of the aligned ligand-set is generated[48]. The occupancy of every ligand atom is mapped onto a grid according to a smooth function ranging from one at the atom center to zero at its solvent accessible surface. An iso-level surface is then generated again, similar to the RSM approach.

1.3 Cytochrome P450

Cytochromes P450 (CYPs) are a superfamily of membrane-bound hemoproteins. They are enzymes, with a heme-iron catalytic site with the iron coordinated via a cysteine residue. CYPs, generally, catalyze the oxidation of a substrate via electron transfer and hydrogen abstraction. CYPs are membrane-anchored proteins, with molecular weights ranging from 45 to 60 kDa, and they contain large, flexible binding pockets. While CYPs are found in a wide variety of species, the human cytochromes are encoded by 57 genes and 33 pseudogenes and are divided into 42 families and subfamilies.[49]

1.3.1 Importance of Cytochrome P450

CYPs metabolize both endogenous and exogenous compounds[22], which leads to their clinical importance: the CYP superfamily is responsible for the metabolism of the majority of pharmaceutical compounds[23]. Particularly important in drug metabolism are CYP1A2, CYP2C9, CYP2C19, CYP2-D6, and CYP3A4[24]. As drug metabolism and elimination are important factors in the drug discovery and development process, and CYPs play a ubiquitous role in those processes, CYP-drug interactions must be kept constantly in mind when developing new pharmaceuticals[25]. Common concerns are metabolic rate, which is a key factor in therapeutic dosage, and the production of toxic metabolites, which can cause the abandonment of otherwise promising drug candidates[26]. These concerns lead to the desire of medicinal chemists for the ability to alter metabolic rate or by-product production by changing the site of ligand metabolism via CYPs[27,28].

1.3.2 Site of Metabolism Prediction

In order to alter the metabolism of a ligand by CYP, one must be able to predict the ligands' sites of metabolism (SoM). Computational tools have become widely used for the prediction of SoM of CYP substrates[27]. As mentioned previously, computational methods are divided into structure-based, ligand-based, and combined methods, and there exists CYP SoM prediction tools that fall into all of these categories[29,30].

1.3.2.1 Ligand-based Site of Metabolism Prediction

Ligand-based techniques, as mentioned previously, analyze ligands' physicochemical properties to predict the most likely site of metabolism. Such methods include quantum chemical calculation-based reactivity prediction, such as SmartCYP[31], pharmacophore models, rule-based methods, and fingerprint methods[32]. While highly efficient, these methods also ignore important considerations, namely the binding pose of the ligand with its target CYP, as the most energetically favorable metabolic site may not be located in proximity to the catalytic heme.

1.3.2.2 Structure-based Site of Metabolism Prediction

The converse to ligand-based techniques, structure-based techniques calculate interactions between a ligand and a structural model of the CYP enzyme to determine the likely SoM. Structure-based techniques include ligand docking and molecular dynamics simulations[33]. These techniques attempt to predict if a ligand will bind a specific CYP enzyme, and if so, attempt to determine the binding pose of the ligand and which ligand atoms are in close proximity to the catalytic heme. These techniques can be time consuming, especially molecular dynamics simulations, and are highly dependent on the accuracy of the scoring function or force field used in the simulations and require protein structure models, which can be difficult to obtain, especially for membrane-bound proteins such as CYPs.

1.3.2.3. Combined Ligand-based and Structure-based Site of Metabolism Prediction

As mentioned previously, both ligand-based and structure-based methods have significant weakness when predicting SoMs. This has led many groups[34-38], including the Lill group[39], to attempt to combine both types of methods. These combined methods are designed to utilize ligand-based information while being guided by structural constraints. These methods are especially useful when multiple metabolic pathways exist for a compound. An exemplar of this situation is the compound Terbinafine, which is metabolized by at least seven different CYP isozymes and results in multiple different metabolites[40]. The complex interactions between reactivity and ligand-binding pose are difficult to predict using structure-based or ligand-based information only.

The approach previously developed by the Lill group combined the NAT reactivity model developed by Olsen *et al.*[41] with ensemble docking. In ensemble docking, molecular dynamics simulations are performed on a protein crystal structure to produce a diverse ensemble of protein structures. The ligands of interest are then docked to every member of the protein ensemble instead of a single crystal structure. This allows for protein flexibility to generate a more diverse set of ligand binding poses. In our previous method, instead of purely relying on the docking scoring function to determine the best scoring pose, and therefore the predicted site of metabolism as determined by proximity to the catalytic site, the NAT model was included as an additional scoring factor. This skewed the results towards those poses with a reactively favored atom close to the metabolic heme, and produced better predictive results than the NAT model or ensemble docking in isolation.

1.4 Research Summary

The overall goal of my research is the application and development of the advancement of combined ligand-based and structure-based techniques, namely pseudoreceptor-based methods, with a focus on surface-based pseudoreceptors. While the goal of pseudoreceptor methods is to produce a protein-like structure to interact with ligands, there has been a lack of use of protein structural data in the guiding of the creation of the pseudoreceptors. In Chapter 2, analysis of the interaction surface between protein crystal structure and co-crystallized ligand for the refined set of the PDBbind database[42,43] will be presented. These surfaces represent the ideal pseudoreceptor, as they map the true interactions of protein and ligand, and the analysis will show that the majority of protein-ligand interactions can be mapped by a few of Gaussian-based descriptors that have parameters that fall into a small range of values. In Chapter 3, a means of tuning surface-based pseudoreceptors to accurately replicate protein binding pocket topology as from known binding ligands will be presented.

In Chapter 4, I will discuss the implementation of the refinement of our group's previous work on SoM prediction, which includes the use of a modified version of the RAPTOR pseudoreceptor package. The modification was the inclusion of reactivity scores from the SMARTCYP package as term in the RAPTOR scoring function. The motivation for the inclusion of RAPTOR was as a means of generating a model which could reliably select binding poses with the known SoM close to the heme of CYP. This method was implemented as a means to counteract the difficulties arising from the large number of poses generated by the ensemble docking process. The initial modeling was performed on CYP2C9, but was later extended to eight other CYP isozymes.

***Note: Portions of this section previously published in the following papers:

Wilson, GL.; Lill, MA. Integrating structure-based and ligand-based approaches for computational drug design, *Future Medicinal Chemistry*, **2011**, *3*, 735-770.

Wilson, GL.; Lill, MA. Towards a realistic representation in surface-based pseudoreceptor modelling: a PDB-wide analysis of binding pockets, *Molecular Informatics*, **2012,** *31*, 259-271

Kingsley, LJ.; Wilson, GL.; Essex, ME.; Lill, MA. Combining Structure- and Ligand-Based Approaches to Improve Site of Metabolism Prediction in CYP2C9 Substrates. *Pharm. Res.*, **2015**, *32*, 986-1001.

List of References

1. Nicolotti, O; Miscioscia, TF; Carotti, A; Leonetti, F; and Carotti, A. An integrated approach to ligand- and structure-based drug design: Development and application to a series of serine protease inhibitors. *J. Chem. Inf. Model.* 48[6], 1211-1226 (2008).
2. Lin, TW; Melgar, MM; Kurth, D et al. Structure-based inhibitor design of AccD5, an essential acyl-CoA carboxylase carboxyltransferase domain of Mycobacterium tuberculosis. *Proc.Natl.Acad.Sci.U.S.A.* 103[9], 3072-3077 (2006).
3. Goodford, PJ. A Computational-Procedure for Determining Energetically Favorable Binding-Sites on Biologically Important Macromolecules. *J. Med. Chem.* 28[7], 849-857 (1985).
4. Goodford, P. Multivariate characterization of molecules for QSAR analysis. *Journal of Chemometrics* 10[2], 107-117 (1996).
5. Nilsson, J; Wikstrom, H; Smilde, A et al. GRID/GOLPE 3D quantitative structure-activity relationship study on a set of benzamides and naphthamides, with affinity for the dopamine D-3 receptor subtype. *J. Med. Chem.* 40[6], 833-840 (1997).
6. Baroni, M; Costantino, G; Cruciani, G et al. Generating Optimal Linear Pls Estimations (Golpe) - An Advanced Chemometric Tool for Handling 3D-Qsar Problems. *Quantitative Structure-Activity Relationships* 12[1], 9-20 (1993).
7. Tanrikulu, Y and Schneider, G. Pseudoreceptor models in drug design. *Nat. Rev. Drug Discov.* 7[8], 667-677 (2008).
8. Cramer, RD; Patterson, DE; and Bunce, JD. Comparative Molecular-Field Analysis (Comfa) .1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J.Am.Chem.Soc.* 110[18], 5959-5967 (1988).
9. Klebe, G. Comparative molecular similarity indices analysis: CoMSIA. *Perspectives in Drug Discovery and Design* 12[87-104 (1998).
10. Klebe, G; Abraham, U; and Mietzner, T. Molecular Similarity Indexes in A Comparative-Analysis (Comsia) of Drug Molecules to Correlate and Predict Their Biological-Activity. *J. Med. Chem.* 37[24], 4130-4146 (1994).
11. Momany, F; Pitha, R; Klimkovsky, VJ; and Venkatchalam, CM. *Expert Systems and Applications in Chemistry*. American Chemical Society, Washington D.C. (1989).
12. Lemmen, C; Lengauer, T; and Klebe, G. FLEXS: A method for fast flexible ligand superposition. *J. Med. Chem.* 41[23], 4502-4520 (1998).
13. Lill, MA and Vedani, A. Combining 4D pharmacophore generation and multidimensional QSAR: Modeling ligand binding to the bradykinin B-2 receptor. *J. Chem. Inf. Model.* 46[5], 2135-2145 (2006).
14. Lemmen, C and Lengauer, T. Computational methods for the structural alignment of molecules. *J.Comput.Aided Mol.Des.* 14[3], 215-232 (2000).
15. Labute, P; Williams, C; Feher, M; Sourial, E; and Schmidt, JM. Flexible alignment of small molecules. *J. of Med. Chem.* 44[10], 1483-1490 (2001).
16. Jewell, NE; Turner, DB; Willett, P; and Sexton, GJ. Automatic generation of alignments for 3D QSAR analyses. *J. Mol. Graph. Model.* 20[2], 111-121 (2001).

17. Kramer, A; Horn, HW; and Rice, JE. Fast 3D molecular superposition and similarity search in databases of flexible molecules. *J.Comput.Aided Mol.Des.* 17[1], 13-38 (2003).

18. Korhonen, SP; Tuppurainen, K; Laatikainen, R; and Perakyla, M. FLUFF-BALL, a template-based grid-independent superposition and QSAR technique: Validation using a benchmark steroid data set. *J.Chem.Inf.Comput.Sci.* 43[6], 1780-1793 (2003).

19. Girones, X and Carbo-Dorca, R. TGSA-flex: Extending the capabilities of the topo-geometrical superposition algorithm to handle flexible molecules. *J.Comput.Chem.* 25[2], 153-159 (2004).

20. Ronkko, T; Tervo, AJ; Parkkinen, J; and Poso, A. BRUTUS: Optimization of a grid-based similarity function for rigid-body molecular superposition. II. Description and characterization. *J.Comput.Aided Mol.Des.* 20[4], 227-236 (2006).

21. Tervo, AJ; Ronkko, T; Nyronen, TH; and Poso, A. BRUTUS: Optimization of a grid-based similarity function for rigid-body molecular superposition. 1. Alignment and virtual screening applications *J. Med. Chem.* 48[12], 4076-4086 (2005).

22. Guengerich FP. Cytochrome p450 enzymes in the generation of commercial products. *Nat. Rev. Drug Discov*. 1(5), 359-66 2002.

23. Nebert, DW; Russell, DW. Clinical importance of the cytochromes P450. *The Lancet*. 360[9340], 1107-1182 (2002).

24. Wienkers LC, Heath TG. Predicting in vivo drug interactions from in vitro drug discovery data. *Nat. Rev. Drug Discov*. 4(10), 825-33 2005.

25. Emoto C, Murase S, Iwasaki K. Approach to the prediction of the contribution of major cytochrome P450 enzymes to drug metabolism in the early drug-discovery stage. *Xenobiotica*. 36(8), 671-83 2006.

26. Thompson RA, Isin EM, Li Y, Weaver R, Weidolf L, Wilson I, et al. Risk assessment and mitigation strategies for reactive metabolites in drug discovery and development. *Chem. Biol. Interact.* 192(1–2), 65-71 2011

27. Trunzer M, Faller B, Zimmerlin A. Metabolic Soft Spot Identification and Compound Optimization in Early Discovery Phases Using MetaSite and LC-MS/MS Validation. *J. Med. Chem.* 52(2), 329-35 2008.

28. Kirkpatrick P. Drug metabolism: Seeking the soft spots. *Nat. Rev. Drug Discov.* 8(3), 196- 2009

29. Crivori P, Poggesi I. Computational approaches for predicting CYP-related metabolism properties in the screening of new drugs. *Eur. J. Med. Chem.* 41(7), 795-808 2006.

30. Kirchmair J, Williamson MJ, et al. Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms. *J. Chem. Inf. Model.* 52(3), 617-48 2012.

31. Rydberg P, Gloriam DE, Zaretzki J, Breneman C, Olsen L. SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. *ACS Med. Chem. Lett.* 1(3), 96-100 2010.

32. Cruciani G, Carosati E, De Boeck B, Ethirajulu K, Mackie C, Howe T, et al. MetaSite: Understanding Metabolism in Human Cytochromes from the Perspective of the Chemist. *J. Med. Chem.* 48(22), 6970-9 2005.

33. Hritz J, de Ruiter A, Oostenbrink C. Impact of Plasticity and Flexibility on Docking Results for Cytochrome P450 2D6: A Combined Approach of Molecular Dynamics and Ligand Docking. *J. Med. Chem.* 51(23), 7469-77 2008.

34. Moors SLC, Vos AM, Cummings MD, *et al*. Structure-Based Site of Metabolism Prediction for Cytochrome P450 2D6. *J. Med. Chem.* 54(17), 6098-105 2011.

35. Zaretzki J, Bergeron C, Rydberg P, Huang TW, Bennett KP, Breneman CM. RS-Predictor: A New Tool for Predicting Sites of Cytochrome P450-Mediated Metabolism Applied to CYP 3A4. *J. Chem. Inf. Model.* 51(7), 1667-89 2011.

36. Tyzack JD, Williamson MJ, Torella R, Glen RC. Prediction of Cytochrome P450 Xenobiotic Metabolism. *J. Chem. Inf. Model.* 53(6), 1294-305 2013.

37. Campagna-Slater V, Pottel J, Therrien E, Cantin L-D, Moitessier N. Development of a Computational Tool to Rival Experts in the Prediction of Sites of Metabolism of Xenobiotics by P450s. J. *Chem. Inf. Model.* 52(9), 2471-83 2012.

38. Li J, Schneebeli ST, Bylund J, Farid R, Friesner RA. IDSite: An Accurate Approach to Predict P450-Mediated Drug Metabolism. *J. Chem. Theo. Comp.* 7(11), 3829-45 2011.

39. Danielson ML, Desai PV, Mohutsky MA, Wrighton SA, Lill MA. Potentially increasing the metabolic stability of drug candidates via computational site of metabolism prediction by CYP2C9. *Eur. J. Med. Chem.* 46(9), 3953-63 2011.

40. Vickers AEM, Sinclair JR, Zollinger M, Heitz F, Glänzel U, Johanson L, et al. Multiple Cytochrome P-450s Involved in the Metabolism of Terbinafine Suggest a Limited Potential for Drug-Drug Interactions. *Drug Metab. Dispos*. 27(9), 1029-38 1999.

41. Rydberg P, Vasanthanathan P, Oostenbrink C, Olsen L. Fast Prediction of Cytochrome P450 Mediated Drug Metabolism. *ChemMedChem*. 4(12):2070-9 2009.

42. Wang, R, Fang, X, Lu, Y, Yang, CY, Wang, S. The PDBbind Database: Methodologies and updates, *J. Med. Chem.* 48(12), 4111-4119 2005.

43. Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* 47(12), 2977-2980 2004.

44. Wilson G, Lill M. Integrating structure-based and ligand-based approaches for computational drug design. *Fut Med. Chem.* 3(6) ,735–750 2011.

45. Hahn M. Receptor surface models. 1. Definition and construction. *J. Med. Chem* 38(12), 2080–90 1995.

46. Hahn M, Rogers D. Receptor surface models. 2. Application to quantitative structure-activity relationships studies. *.J Med. Chem.* 38(12), 2091–102 1995.

47. Hahn M, Rogers D. Receptor surface models. *Perspect. Drug Discov.* 12(14), 117–133 1998.

48. Lill MA, Vedani A, Dobler M. Raptor: combining dual-shell representation, induced-fit simulation, and hydrophobicity scoring in receptor modeling. *J. Med. Chem.* 47(25), 6174–6186 2004.

49. Nebert DW, Nelson DR, *et al.* The P450 Superfamily. *DNA*. 8(1), 1-13 1989.

50. Kato Y, Itai A, and Iitaka, Y. A Novel Method for Superimposing Molecules and Receptor Mapping. *Tetrahedron* 43(22), 5229-5236 1987.

51. Kato Y, Inoue A, Yamada M, Tomioka N, and Itai A. Automatic Superposition of Drug Molecules Based on Their Common Receptor-Site. *J.Comput.Aided Mol.Des.* 6(5), 475-486 1992.

52. Guccione S, Doweyko, AM, Chen, H, Barretta, GU, and Balzano, F. 3D-QSAR using "Multiconformer" alignment: The use of HASL in the analysis of 5-HT$_{1A}$ thienopyrimidinone ligands. *J.Comput.Aided Mol.Des.* 14(7), 647-657 2000.
53. Andrews PR, Quint G, Winkler DA et al. Morpheus - A Conformation-Activity Relationships and Receptor Modeling Package. *J.Mol.Graph.* 7(3), 138-145 1989.
54. Lloyd DG, Buenemann CL, Todorov NP, Manallack DT, and Dean PM. Scaffold Hopping in De Novo Design. Ligand Generation in the Absence of Receptor Information. *J. Med. Chem.* 47(3), 493-496 2004.
55. Todorov NP and Dean PM. A branch-and-bound method for optimal atom-type assignment in de novo ligand design. *J.Comput.Aided Mol.Des.* 12(4), 335-349 1998.
56. Crippen GM. Validation of EGSITE2, a mixed integer program for deducing objective site models from experimental binding data. *J. Med. Chem.* 40(20), 3161-3172 1997.
57. Crippen GM. Voronoi Binding-Site Models. *J.Comput.Chem.* 8(7), 943-955 1987.
58. Zbinden P, Dobler M, Folkers G, and Vedani A. PrGen: Pseudoreceptor modeling using receptor-mediated ligand alignment and pharmacophore equilibration. *QSAR.* 17(2), 122-130 1998.
59. Vedani A, Zbinden P, Snyder JP, and Greenidge PA. Pseudoreceptor Modeling - the Construction of 3-Dimensional Receptor Surrogates. *J.Am.Chem.Soc.* 117(17), 4987-4994 1995.
60. Vedani A, Dobler M, and Zbinden P. Quasi-atomistic receptor surface models: A bridge between 3-D QSAR and receptor modeling. *J.Am.Chem.Soc.* 120(18), 4471-4477 1998.
61. Vedani A, Zbinden P, and Snyder JP. Pseudo-Receptor Modeling - A New Concept for the 3-Dimensional Construction of Receptor-Binding Sites. *J.Recept.Res.* 13(1-4), 163-177 1993.
62. Vedani A and Dobler M. 5D-QSAR: The key for simulating induced fit? *J. Med. Chem.* 45(11), 2139–2149 2002.
63. Vedani A, Dobler M, and Lill MA. Combining protein modeling and 6D-QSAR – Simulating the binding of structurally diverse ligands to the estrogen receptor. *J. Med. Chem.* 48(11), 3700–3703 2005.
64. Walters DE and Hinds RM. Genetically Evolved Receptor Models - A Computational Approach to Construction of Receptor Models. *J. Med. Chem.* 37(16), 2527-2536 1994.
65. Pei J, Zhou JJ, Xie GR, Chen HM, and He X. PARM: A practical utility for drug design. *J. Mol. Graph. Model.* 19(5), 448-454 2001.
66. Chen HM, Zhou JJ, and Xie GR. PARM: A genetic evolved algorithm to predict bioactivity. *J.Chem.Inf.Comput.Sci.* 38(2), 243-250 1998.
67. Pei JF and Zhou JJ. Flexible atom receptor model. *Acta Chimica Sinica* 60(6), 973-979 2002.
68. Chae CH, Yoo SE, and Shin W. Novel receptor surface approach for 3D-QSAR: The weighted probe interaction energy method. *J.Chem.Inf.Comput.Sci.* 44(5), 1774-1787 2004.
69. Pei JF, Chen H, Liu ZM et al. Improving the quality of 3D-QSAR by using flexible-ligand receptor models. *J. Chem. Inf. Model.* 45(6), 1920-1933 2005.

70. Chen W and Gilson MK. ConCept: De novo design of synthetic receptors for targeted ligands. *J. Chem. Inf. Model.* 47(2), 425-434 2007.
71. McMartin C and Bohacek RS. QXP: Powerful, rapid computer algorithms for structure-based drug design. *J.Comput.Aided Mol.Des.* 11(4), 333-344 1997.
72. Hay BP and Firman TK. HostDesigner: A program for the de novo structure-based design of molecular receptors with binding sites that complement metal ion guests. *Inorg.Chem.* 41(21), 5502-5512 2002.
73. Wolber G and Langer T. LigandScout: 3-d pharmacophores derived from protein-bound Ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* 45(1), 160-169 2005.
74. Wolber G and Langer T. LigandScout: Interactive automated pharmacophore model generation from ligand-target complexes. *Abstracts of Papers of the American Chemical Society* 229, U611 2005.
75. Sato T, Honma T, and Yokoyama S. Combining Machine Learning and Pharmacophore-Based Interaction Fingerprint for in Silico Screening. *J. Chem. Inf. Model.* 50(1), 170-185 2010.
76. Baroni M, Cruciani G, Sciabola S, Perruccio F, and Mason JS. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): Theory and application. *J. Chem. Inf. Model.* 47(2), 279-294 2007.
77. Mason JS, Morize I, Menard PR et al. New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* 42(17), 3251-3264 1999.
78. Weill N and Rognan D. Development and Validation of a Novel Protein-Ligand Fingerprint To Mine Chemogenomic Space: Application to G Protein-Coupled Receptors and Their Ligands. *J. Chem. Inf. Model.* 49(4), 1049-1062 2009.
79. Deng Z, Chuaqui C, and Singh J. Structural interaction fingerprint (SIFt): A novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* 47(2), 337-344 2004.
80. Chuaqui C, Deng Z, and Singh J. Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. *J. Med. Chem.* 48(1), 121-133 2005.
81. Kelly MD and Mancera RL. Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. *J.Chem.Inf.Comput.Sci.* 44(6), 1942-1951 2004.
82. Perez-Nueno VI, Rabal O, Borrell JI, and Teixido J. APIF: A New Interaction Fingerprint Based on Atom Pairs and Its Application to Virtual Screening. *J. Chem. Inf. Model.* 49(5), 1245-1260 2009.
83. Deng Z, Chuaqui C, and Singh J. Knowledge-based design of target-focused libraries using protein-ligand interaction constraints. *J. Med. Chem.* 49(2), 490-500 2006.
84. Nandigam RK, Kim S, Singh J, and Chuaqui C. Position Specific Interaction Dependent Scoring Technique for Virtual Screening Based on Weighted Protein-Ligand Interaction Fingerprint Profiles. *J. Chem. Inf. Model.* 49(5), 1185-1192 2009.
85. Marcou G and Rognan D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* 47(1), 195-207 2007.

86. Venhorst J, Nunez S, Terpstra JW, and Kruse CG. Assessment of scaffold hopping efficiency by use of molecular interaction fingerprints. *J. Med. Chem.* 51(11), 3222-3229 2008.
87. Mpamhanga CP, Chen BN, Mclay IM, and Willett P. Knowledge-based interaction fingerprint scoring: A simple method for improving the effectiveness of fast scoring functions. *J. Chem. Inf. Model.* 46(2), 686-698 2006.
88. Crisman TJ, Sisay MT, and Bajorath J. Ligand-Target Interaction-Based Weighting of Substructures for Virtual Screening. *J. Chem. Inf. Model.* 48(10), 1955-1964 2008.
89. Tan L, Lounkine E, and Bajorath J. Similarity Searching Using Fingerprints of Molecular Fragments Involved in Protein-Ligand Interactions. *J. Chem. Inf. Model.* 48(12), 2308-2312 2008.
90. Tan L and Bajorath J. Utilizing Target-Ligand Interaction Information in Fingerprint Searching for Ligands of Related Targets. *Chemical Biology & Drug Design* 74(1), 25-32 2009.
91. Tan L, Vogt M, and Bajorath J. Three-Dimensional Protein-Ligand Interaction Scaling of Two-Dimensional Fingerprints. *Chemical Biology & Drug Design* 74(5), 449-456 2009.
92. McGregor MJ and Pallai PV. Clustering of large databases of compounds: Using the MDL "keys" as structural descriptors. *J.Chem.Inf.Comput.Sci.* 37(3), 443-448 1997.
93. Batista J, Tan L, and Bajorath J. Atom-Centered Interacting Fragments and Similarity Search Applications. *J. Chem. Inf. Model.* 50(1), 79-86 2010.
94. Kroemer RT, Vulpetti A, McDonald JJ et al. Assessment of docking poses: Interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J.Chem.Inf.Comput.Sci.* 44(3), 871-881 2004.
95. Naumann T and Matter H. Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: Target family landscapes. *J. Med. Chem.* 45(12), 2366-2378 2002.
96. MACCS structural keys, MDL Elsevier, San Leandro, CA, USA. http://www.mdl.com 2002
97. Willett P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* 11(23-24), 1046-1053 2006.
98. Jain AN. Virtual screening in lead discovery and optimization. *Current Opinion in Drug Discovery & Development* 7(4), 396-403 2004.
99. Kitchen DB, Decornez H, Furr JR, and Bajorath J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery* 3(11), 935-949 2004.
100. Shoichet BK. Virtual screening of chemical libraries. *Nature* 432(7019), 862-865 2004.
101. Tan L, Geppert H, Sisay MT, Gutschow M, and Bajorath J. Integrating Structure- and Ligand-Based Virtual Screening: Comparison of Individual, Parallel, and Fused Molecular Docking and Similarity Search Calculations on Multiple Targets. *Chemmedchem* 3(10), 1566-1571 2008.
102. Fukunishi Y. Structural ensemble in computational drug screening. *Expert Opinion on Drug Metabolism & Toxicology* 6(7), 835-849 2010.

103. Fukunishi Y and Nakamura H. Prediction of protein-ligand complex structure by docking software guided by other complex structures. *J. Mol. Graph. Model.* 26(6), 1030-1033 2008.

104. Hirokawa T. Receptor-ligand docking simulation for membrane proteins. *Yakugaku Zasshi-Journal of the Pharmaceutical Society of Japan* 127(1), 123-131 2007.

CHAPTER 2. AN ANALYSIS OF BINDNG POCKETS

2.1 Overview

Pseudoreceptor models are intended to contain key protein–ligand interactions, and to map the appropriate spatial information content of these interactions. The aim of pseudoreceptor modelling is to generate surrogates of the 3-D structure of the protein binding site that can be used for structure-based drug design applications such as virtual screening, rationally modifying or proposing new small molecules complementary to the pseudoreceptor model, and predicting binding affinities of potential ligands. Although several types of pseudoreceptor representations exist[1,2], one popular class are surface-based pseudoreceptor models that represent the binding site of the target protein by selected surfaces. Of particular interest is the solvent-accessible surface, as it represents the 3-D-space most critical for the complementary contacts between protein and ligand. Hydrogen bonds and van der Waals interactions are particularly strong at the protein-ligand interface. Thus, these surfaces can provide a rather complete representation of the protein-ligand contacts while reducing the number of descriptors compared to grid-based approaches.

The method RAPTOR[11] is an example of a surface- based pseudoreceptor approach. Additionally, RAPTOR accounts for both ligand and protein flexibility. In general the RAPTOR algorithm works by distributing hydrophobic and hydrogen bond properties representing the surrogate of the target protein onto a surface surrounding an aligned set of ligand molecules until the interaction between these surface properties and the ligands reproduces the experimental binding affinities of the compounds. A scoring function is then utilized to measure the interaction strength between surface properties and ligand atoms.

The critical question is how the different physicochemical properties are distributed onto the pseudoreceptor surface. In methods such as RAPTOR the surface is typically represented by several hundred points. In the most naïve approach those points are treated independently from each other, and overfitting may occur during optimization of the pseudoreceptor model. To reduce the number of descriptors in RAPTOR, we defined patches of surface points that were empirically forced to adopt similar physicochemical properties. The patch size and the transition between patches, however, are user-defined and may not reflect accurately the distribution of physico-chemical properties in experimental protein structures.

In this chapter, we address the question if the physico-chemical properties on the solvent-accessible surface of experimentally determined protein structures can be accurately modelled by a small number of surface descriptors. First, we analysed binding pocket surfaces of a large set of experimentally determined protein-ligand complexes and used 2-D Gaussian functions to fit the surface properties. The fitted property values differ from the original values on average by 15-25%, and on average six Gaussian functions are necessary to model each surface property. These descriptors will allow for a more realistic pseudoreceptor representation of the binding site compared to our current empirical patching model implemented in RAPTOR and limit the number of descriptors, thereby reducing the potential of overfitting throughout the QSAR optimization phase.

2.2 Database Preparation

For our analysis of protein-ligand interaction surfaces the PDBBind Database[3,4] was chosen, as it contains target protein structures co-crystallized with small molecule ligands. The refined set of the PDBBind database, a set of protein-small molecule crystal complexes manually reviewed for resolution, binding affinity data, protein amino acid composition, and ligand molecular and common element composition criteria, was prepared as input data to our Protein-Ligand Surface Interaction Analysis (PLSIA) program (see next section). The REDUCE program[5] was utilized to add missing atoms to the PDB structures and to optimize the protein's hydrogen-bond network by adjusting Asn, Gln and His side-chain orientations, as well as the tautomeric and protonation state of His residues. AMBER atom types and partial charges were then assigned to the optimized protein structures using the AMBER03[6] force field parameter file. Parameters were not assigned to the ligand structures as they are used solely for the definition of the binding pocket.

2.3 PLSIA Algorithm

2.3.1 Surface Triangulation

The Protein-Ligand Surface Interaction Analysis (PLSIA) program operated by loading each protein structure and generating a separate PDB file in which the ligand has been removed. The pseudo-holo form was used for the calculation of triangulated surfaces on the exterior protein surface and any cavity using the MSRoll program[7]. MSRoll uses a rolling probe method to determine the solvent-accessible surface. The cavity or surface closest to the ligand was identified using a distance calculation between the ligand atoms and the closest surface points and was selected as the protein- ligand interaction surface. In order to produce a surface with a larger number of smaller triangles, the tessellation fitness parameter was set to 0.5 radians, the default settings were used for all additional parameters.

MSRoll generated a triangulated representation of each exterior and cavity surface, however, the triangulation is often heterogeneous, with triangles varying significantly in size. The COALESCE program[8] was used to regularize the triangulated surfaces. COALESCE loads the MSRoll output and combines small triangles, those with edges with less than half the average edge length (which is typically around 0.5 Å), and fixes dangling vertices resulting in a smooth triangulated surface with similar triangle sizes. COALESCE also standardizes the direction of the normal vectors of each triangle, always pointing away from the protein. After this refinement, all triangle vertices more than 8 Å from any ligand atom were removed.

2.3.2 Protein Property Calculation

After identifying and isolating the binding pocket surface, the program PLSIA determined the electrostatic, hydrogen bond, and hydrophobic field of the protein mapped onto each triangle vertex. The results of these equations were parameterized such that an output value of one is approximately one kcal/mol of interaction energy.

2.3.2.1 Electrostatic Calculations

In PLSIA a partial charge was assigned to each protein atom using the Amber03 force field. The Coulombic potential on each surface vertex, *s*, generated by the protein atoms within 12 Å was determined by the following equation:

$$V_{Coul}^{s} = \sum_{i=1}^{n} \frac{q_i}{D(r_{si}) \cdot r_{si}} \tag{2.1}$$

Here, n is the number of protein atoms within 12 Å of the surface point s, q_i is the partial charge of protein atom i, and $D(r_{si})$ is the distance-dependent dielectric, which in this case is r_{si} itself, and r_{si} is the distance between the vertex s and atom i.

2.3.2.2 Hydrogen Bond Calculations

All protein atoms capable of forming hydrogen-bonds were identified in PLSIA. Next, the availability of the donors and acceptors to form protein-ligand hydrogen bonds was determined. Donors and acceptors form intra-protein hydrogen bonds if there was a complimentary hydrogen-bonding partner identified within 2.5Å and a maximum angle of 40 degrees between the donor hydrogen atom, donor heavy atom and the acceptor atom. Any donor or acceptor atom occupied by intra-protein hydrogen bonds was not considered for further calculation of the hydrogen-bond fields on the protein surface. Accessible hydrogen bonding protein atoms that were within 4 Å of a surface vertex were used to calculate the hydrogen bond potential on the vertex point *s* using the following equation:

$$V_{HB}^{s} = \sum_{i=1}^{n} f_{Fermi}\left(r_{si};0.5,2.3\right)\cdot f_{Fermi}\left(\theta;10^{o},50^{o}\right) \tag{2.2a}$$

with

$$f_{Fermi}\left(x;a,b\right) = \frac{1}{1+\exp\left(\frac{1}{a}\left(x-b\right)\right)} \tag{2.2b}$$

Here, r_{si} is the distance between the vertex *s* and the hydrogen bond partner (hydrogen for donor and heteroatom for acceptor), and θ *is* the hydrogen bond angle. For hydrogen bond donors, this angle is between the donor hydrogen atom, the donor heavy atom, and the vertex. For hydrogen bond acceptors, the angle is between the acceptor's lone pair, the acceptor atom, and the vertex. The donor potential and acceptor potential were determined separately, and a value for both was assigned to each vertex. In addition, the protein atoms which provided the strongest contribution to each vertex were identified and stored.

2.3.2.3 Hydrophobic Calculations

PLSIA assigned a partial logP value to each protein atom using the methodology of Wildman and Crippen[9]. The overall hydrophobic field spawned by all protein atoms onto each vertex *s* was computed using the following equation:

$$V_{HO}^{s} = \sum_{i=1}^{n} \log P_i \cdot f_{Fermi}\left(r_{si};1.5,2.5\right) \tag{2.3}$$

Here, n is the number of protein atoms within 3 Å of the vertex s, $\log P_i$ is the partial log P value of protein atom i, and r_{si} is the distance between the vertex s and atom i.

2.3.3 Gaussian Preparation

After the physico-chemical properties of interest have been mapped onto the vertex points of the solvent-accessible surface of the protein, several calculations had to be performed to project the 3-D surface onto a 2-D projection map allowing subsequent 2-D fits using Gaussian network models (see section 2.2.6). First, an analysis of the connectivity of the vertex points was performed to generate the Shortest Path Array (SPA). This array was an NxN matrix, where N is the number of surface vertex points, describing the shortest connectivity between two vertices along the edges of the triangulated surface. The generation of the SPA involved the use of an NxN Edge matrix in which all vertices *i* and *j*, that are connected via an edge of the triangulated surface were assigned a value of one to their corresponding entries $\text{Edge}_{i,j}$ and $\text{Edge}_{j,i}$ in the Edge matrix. All entries $\text{Edge}_{i,j}$ that corresponded to unconnected vertices *i* and *j* were set to zero. A brute-force search along the connected edges between all vertices using the Edge matrix was used to calculate the smallest number of edges separating two vertices, and this value was stored in the SPA. Thus the SPA recorded the smallest separation between two vertices along the edges of the triangulated surface.

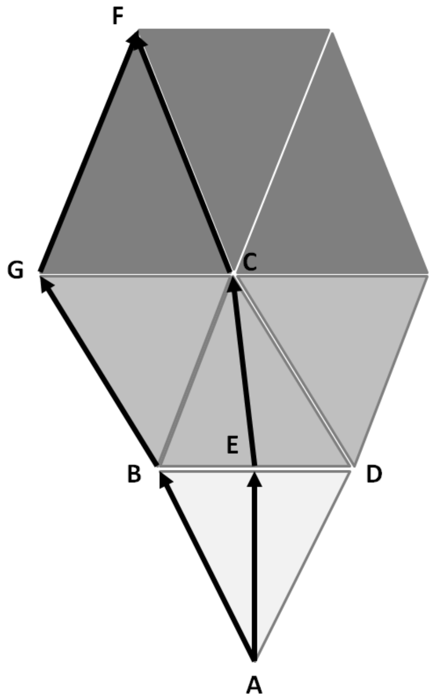


Figure 2.1: Possible paths for vertices separated by three edges. Central path which illustrates the bisecting path algorithm is shorter than left path composed solely of edge traveling.

Next, the relative coordinates of the surface points in 3-D representation of the protein surface were projected into a 2-D representation to allow for a fit with 2-D Gaussian functions: PLSIA approximates the 2-D distance along the interaction surfaces (Figure 2.1). This calculation was trivial for adjacent vertices (Figure 2.1, points A, B), but became more complicated for more separated surface vertex points. For all adjacent vertex points, the direct 3-D distance between the two vertex points was calculated and stored.

For points (Figure 2.1, points A, C) separated by two edges (SPA = 2), the following process was used: First, all intermediate points were identified that are directly connected to the two target vertices (SPA = 1). Due to the triangulation of the surface, some vertices shared two intermediate points (B, D). For those cases, the distance between the two target vertices was defined by finding the point on the intermediate edge point (E) that had the smallest sum of distances to the target vertices (A, C). This distance was determined computing the smallest distance along the triangulated surface between points A and C passing through 100 equally separated points along the edge BD. This represented the shortest path along the triangulated surface between the target vertices (A, C) for a given pair of intermediates (B, D). This distance was computed for all vertices with SPA = 2, and the shortest distance for each pair was stored. For points separated by successively higher SPA values, the distance was determined by finding the shortest distance given a single intermediate point: For vertices with 3 edge separations (SPA = 3; Figure 2.1, points A, F) all intermediates with an SPA = 1 to a target vertex (e.g. CF, GF) and an edge with SPA = 2 (e.g. AC, AG) to the other vertex were compared; for vertices with 4 edge separations of all SPA=3/SPA=1 and SPA=2/SPA=2 pairings were compared and the shortest distance for each target vertex pair was stored. This procedure was then generalized for larger edge separations.

2.3.4 Patching Process

Next, the surface was divided into patches that represent local maxima of the physico-chemical properties that were subsequently fitted using Gaussian functions (see section 2.2.6). This patching was accomplished by the following process: First, dependent on the studied property, the maximum or minimum value of the property on a vertex was identified, and this vertex was defined to be the origin of the first patch. Starting from this origin vertex, the vertices with increasing separation from the origin were examined: First all vertices with SPA=1, then SPA=2, etc. were examined to determine if they are added to the current patch, until an empirically defined maximum edge separation of nine was reached.

In order for a vertex to be included in the current patch, it must have fulfilled a number of conditions: First, the absolute value of the property had to exceed a certain minimum property value, set to 0.1 kcal/mol for all properties, except hydrophobicity which was set to 0.05 kcal/mol. This condition was introduced to limit the patches to those vertices that represent significant magnitudes of the properties of interest. This condition also allowed for the separation of the electrostatic potential into patches with positive and negative values using zero, or more precisely the region between -0.1 and 0.1 kcal/mol as boundary criteria. The second condition was that a vertex must be connected by at least a single edge to a vertex already included in the patch definition. This condition was used to prevent two nearby, but separated patches from being combined into a single patch. The third condition for defining a patch was that it must contain at least four vertices, as that is the minimum number of points required to fit a 2-D Gaussian containing four variables (see section 2.3.6). Additionally, the hydrogen bonding parameters had a fourth patching criterion: all members of a patch had to share the same strongest contributing protein atom as identified in the property calculation process. This condition aided in separating overlapping hydrogen bonding patches caused by different hydrogen-bonding protein groups.

2.3.5 Coordinate Transformation

The final step prior to the Gaussian fit was the transformation of the 3-D coordinates of the surface vertices of a patch to 2-D coordinates (Figure 2.2). The following list details the process of this transformation.

1. The center of the patch was defined as the origin (O)
2. The normal vector of the origin (ON) was calculated by averaging the normal vectors of all triangles of which O is a member.
3. A transformation plane was defined by the plane that passes through O and is perpendicular to ON.
4. All other vertices of the patch were projected into this plane along the vectors normal to the plane (e.g. C → C’ or A → A’).
5. A reference axis for use with 2-D polar coordinates was defined by the vector between origin O and vertex C’ closest to the origin point.

6. Using this reference axis and the normal vector of the plane, angles for each of the projected vertices were determined (Figure 2.2).

7. The surface distance between each vertex A and O was looked up from the data calculated in section 2.2.3.

8. A new point B' was calculated for each A by moving the distance from step 7 along the angle from step 6. (e.g. Point B' has coordinates (Dist_{OA}, $\angle$(A'OC')).

9. Using these polar coordinates each new vertex was translated into 2-D Cartesian coordinates with the origin vertex O having coordinates (0, 0) for use as a reference point (e.g. (Dist_{OA} cos($\angle$(A'OC')), Dist_{OA} sin($\angle$(A'OC'))) for vertex B)'.

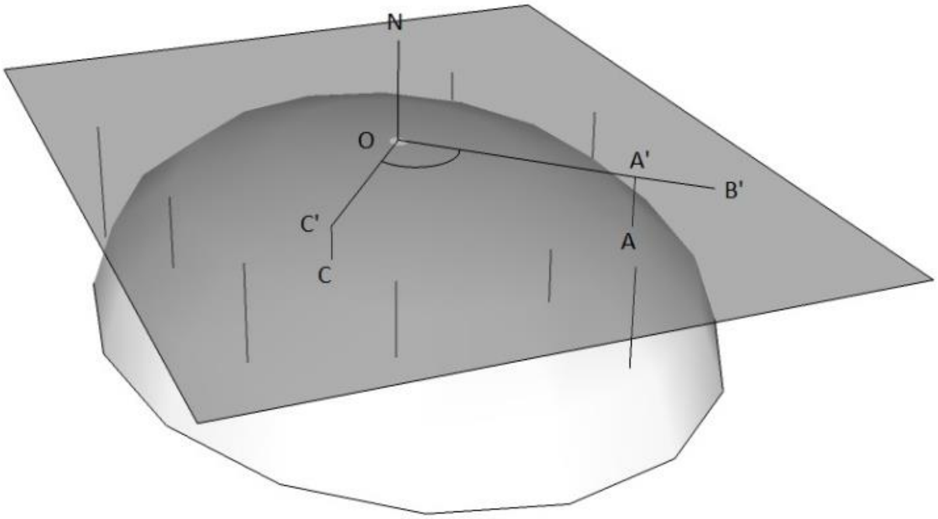


Figure 2.2: Projection of vertices on 3D surface patch into the 2D plane defined by the origin vertex O (maximum or minimum property value) and its normal vector. 3D vertices, e.g. A, are projected into the plane (→A') and are scaled to match the 3D surface distance between O and A (→B').

2.3.6 Gaussian Fitting Process

Once all surface vertices of the patch were translated into 2-D Cartesian coordinate form (x,y), a standard multivariate fitting algorithms was applied to fit a 2-D Gaussian function to each patch:

$$f(x,y) = A\exp\left[-\left(a(x-x_0)^2 + 2b(x-x_0)(y-y_0) + c(y-y_0)^2\right)\right] \tag{2.4}$$

$$a = \frac{\cos^2\theta}{2\sigma_x^2} + \frac{\sin^2\theta}{2\sigma_y^2} \qquad b = -\frac{\sin 2\theta}{4\sigma_x^2} + \frac{\sin 2\theta}{4\sigma_y^2} \qquad c = \frac{\sin^2\theta}{2\sigma_x^2} + \frac{\cos^2\theta}{2\sigma_y^2}$$

with σ_x representing one axis of the Gaussian, σ_y the remaining axis, θ is rotation of the Gaussian axes with respect to the standard Cartesian axes, and *A* is the amplitude of the Gaussian.

PLSIA used the non-linear least-squares fitting algorithm from the GNU Scientific Library[10], based on the Levenberg-Marquadt algorithm, to fit a standard 2-D Gaussian function to each patch. The fitting process was run for 10000 steps and was repeated for up to 81 different initial parameter settings. Each parameter, amplitude, θ, σ_x, and $\sigma_{y,}$ had three possible initial settings: 0.25, 0.75, 1.5. Different permutations of these starting parameters were run until the Gaussian fit converged to a solution with a low sum square error (the average percent error across the patch is less than 10%) or until all 81 initial parameter sets had been evaluated, in which case the run producing the smallest error was selected.

2.3.7 Iteration and Re-patching Process

After fitting the first patch, the patching and fitting process continued according to two different schemes. In the first scheme, called One Pass Fit patching, all vertices of the previously fitted patch were removed from further consideration for subsequent patching and Gaussian fitting. The patching and fitting process was repeated for all remaining vertices until no further vertices meet the patch criteria. In this scheme, each point was fitted to a Gaussian function at most once for any studied physico-chemical property.

The second scheme, called Residual Fit patching, allowed for each vertex to be fitted multiple times. This was done to see if the surface was more accurately represented by one patch per property per surface region or multiple overlapping patches. This occurs by subtracting the physico-chemical values of the Gaussian fit from the original value for every vertex of the current patch resulting in residual values. These residual values were then assigned to each vertex and the modified vertices were further considered in the patching algorithm as viable candidates.

2.3.8 Clique Detection Analysis Using Patch Centers

The Gaussian functions fitted to the patches model the physico-chemical properties of the protein projected onto its surface. We expect similar properties on the surface for the same protein bound to different ligands if no significant conformational change occurs. To test this hypothesis, we performed clique detection between the patch centers for each pair of proteins. The center points of the Gaussian fits produced by the PLSIA program were considered as pharmacophores representing protein properties An edge array for each protein surface was generated storing the distances and center types (e.g. donor-donor, donor-acceptor, etc.) for all pairs of centers.

These arrays were used in a clique detection algorithm performing an exhaustive search to identify the maximum number of patch centres for which all pairwise distances between the two proteins match. A distance between two centers was considered a match if the distances for the two proteins were within a user-defined tolerance of 0.75 Å and if the centers had matching corresponding property types. A score *S* was computed for each pair of proteins *i* and *j* to measure the number of common patch centers in the maximum common clique:

$$S_{ij} = \frac{n_{ij}^{centers}}{\min(n_i^{centers}, n_j^{centers})} \qquad (2.5)$$

where $n_{ij}^{centers}$ is the number of matching centers in the clique and $n_i^{centers}$ is the number of centers for protein *i*. The number of matching centers was normalized by the smaller number of total centers of the two compared protein in order to correct for the variation in binding pocket size due to variation in size of the co-crystallized ligands.

2.4 PLSIA Results

2.4.1 Quality of Fit

We used the following measure for evaluating the quality of the fits of the PLSIA algorithm:

$$E_p = \frac{\sum_{i=1}^{n} \left| F_i - C_i \right|}{\sum_{i=1}^{n} C_i} \tag{2.6}$$

E_p represents the average relative deviation over n evaluated vertices for a property p between the fitted values F_i at vertex i and computed initial surface value C_i at vertex i. We used this criterion to evaluate two different analysis schemes using PLSIA, One Pass and Residual Fit, as described under Section 2.3.7.

Table 2.1 displays the average relative error of Gaussian fitting for electrostatic, hydrophobic and hydrogen properties. For electrostatic and hydrophobic properties, the average error is approximately 15% of the initial surface values, however larger errors are observed for donor and acceptor properties. This suggests that the patches can be represented by single Gaussians. We propose that this difference in fitting accuracy between hydrogen-bond and the other properties is at least partially due to the directionality of the hydrogen bond (Figure 2.5). If the hydrogen bond is oriented along the normal vector of the surface, the resulting patch is well characterized by a symmetric function such as the Gaussian function, as with increasing surface distance from the center of the patch both distance and angle increase reducing the interaction potential according to equation 2.2. However, if the hydrogen bond direction is tilted with respect to the surface normal (Figure 2.5), the distance and angle term in equation 2.2 display different maxima on the surface and the resulting interaction potential on the surface will become asymmetric.

Table 2.1: Average Relative Error of Gaussian Fitting for Individual Properties

Algorithm	El.st.	H.phobic	Donor	Acceptor
One Pass Fit	0.159	0.156	0.254	0.225
Residual Fit	0.158	0.144	0.238	0.210

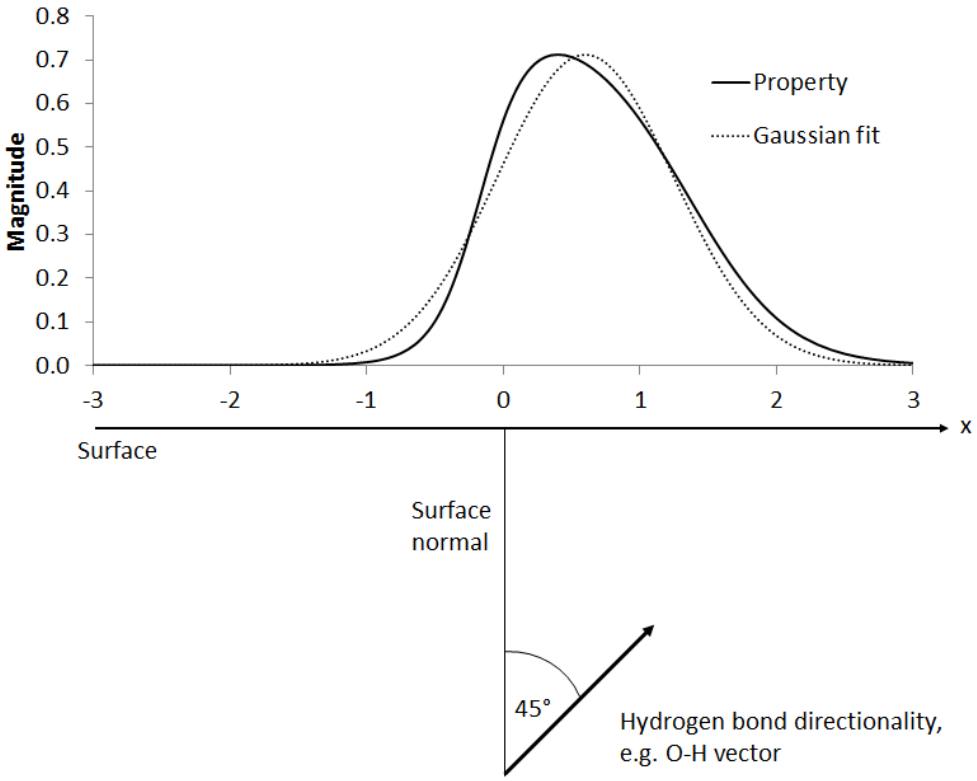


Figure 2.3: Hydrogen-bond interaction potential on the surface using equation 2 together with the optimal Gaussian fit to that distribution for a single hydrogen bond group tilted with respect to the surface normal by 45 degrees. The hydrogen bond group is located two units below the axis at x=0. The angle causes the potential function to have maximum not located at the coordinate center (x=0). The error using equation 5 is approximately 10%.

We also observed that for all properties there is only a small difference in average relative errors between the One Pass and Residual fits. Additionally, we investigated the effect of two central variables in the program, the minimum and maximum sizes of the fitted surface patch, onto the observed error in fit. The absolute minimum size of a patch is four points, as that is the number of unknown variables in the Gaussian function. Changing the minimum size (Table 2.2) showed significant variation in the error of the fits. Some of the error is due to several small patches that are being less well fit by Gaussian functions when a larger patch size is enforced. Consequently, we chose a minimum size of four points for subsequent analysis.

The maximum size of the property patch is governed by the maximum number of edge lengths measured from the surface point representing the maximum value of a property of a patch. We investigated maximum edge lengths of five to nine edges using the Residual Fit method. Comparing the results across all properties, no significant variation in observed error can be noted, as shown in Table 2.3. The aim of this study was to investigate if a small number of surface descriptors can be used to characterize the distribution of physico-chemical properties of the protein. Thus, we decided to allow patches with maximum edge length of nine, as this would prevent the splitting of large patches into smaller fractions, resulting in a decrease in number of surface descriptors.

Table 2.2. Average Relative Error of Gaussian Fitting for Individual Properties with Variation of Minimum Patch Size (Residual Fit). Maximum edge length was set to eight.

Type of interaction	**Minimum Points**		
	4	8	12
El.st.	0.158	0.166	0.169
H.phobic	0.144	0.166	0.185
Donor	0.238	0.270	0.299
Acceptor	0.210	0.231	0.242

Table 2.3: Average Relative Error of Gaussian Fitting for Individual Properties with Variation of Maximum Patch Size (Residual Fit). Minimum number of points in patch was set to four.

Type of interaction	**Maximum edge length**				
	5	6	7	8	**9**
El.st.	0.141	0.148	0.153	0.158	0.158
H.phobic	0.138	0.141	0.141	0.143	0.144
Donor	0.240	0.240	0.238	0.238	0.238
Acceptor	0.210	0.209	0.209	0.209	0.210

2.4.2 Characterization of distributions of properties

We characterized how the different physico-chemical properties are distributed on the surface by analysing the size of the patches, and the magnitude and widths of the fitted Gaussian functions. Here we considered the results from the Residual Fit analysis.

Figure 2.4 shows that the largest portion of patches are small in size, with a maximum number of patches with a size between four and ten (~1-3 Å^2) for all properties except hydrogen bond donors which adopts a broad maximum at a patch size containing about 20 surface points (~5 Å^2). The frequency of obtaining large patches decreases rapidly with size for electrostatic and hydrogen bond properties, however this trend is weaker for hydrophobic properties. Compared to hydrogen bond properties, the interaction potential for hydrophobic contacts (equation 2.3) has a longer interaction range. Furthermore, the hydrophobic function doesn’t contain any directionality information, thus the hydrogen bond patches are more localized and consequently smaller in size compared to the hydrophobic patches. Electrostatic interactions are dependent on the partial charges, which for formally neutral chemical groups are due to differences in electronegativity of bonded atoms. The connected atoms in those groups typically have alternating signs of partial charge. On the contrary, hydrophobic moieties in the binding site of protein consist of a collection of connected atoms. Thus, it is not surprising that hydrophobic surface patches are larger in dimension than electrostatic patches from neutral chemical entities.

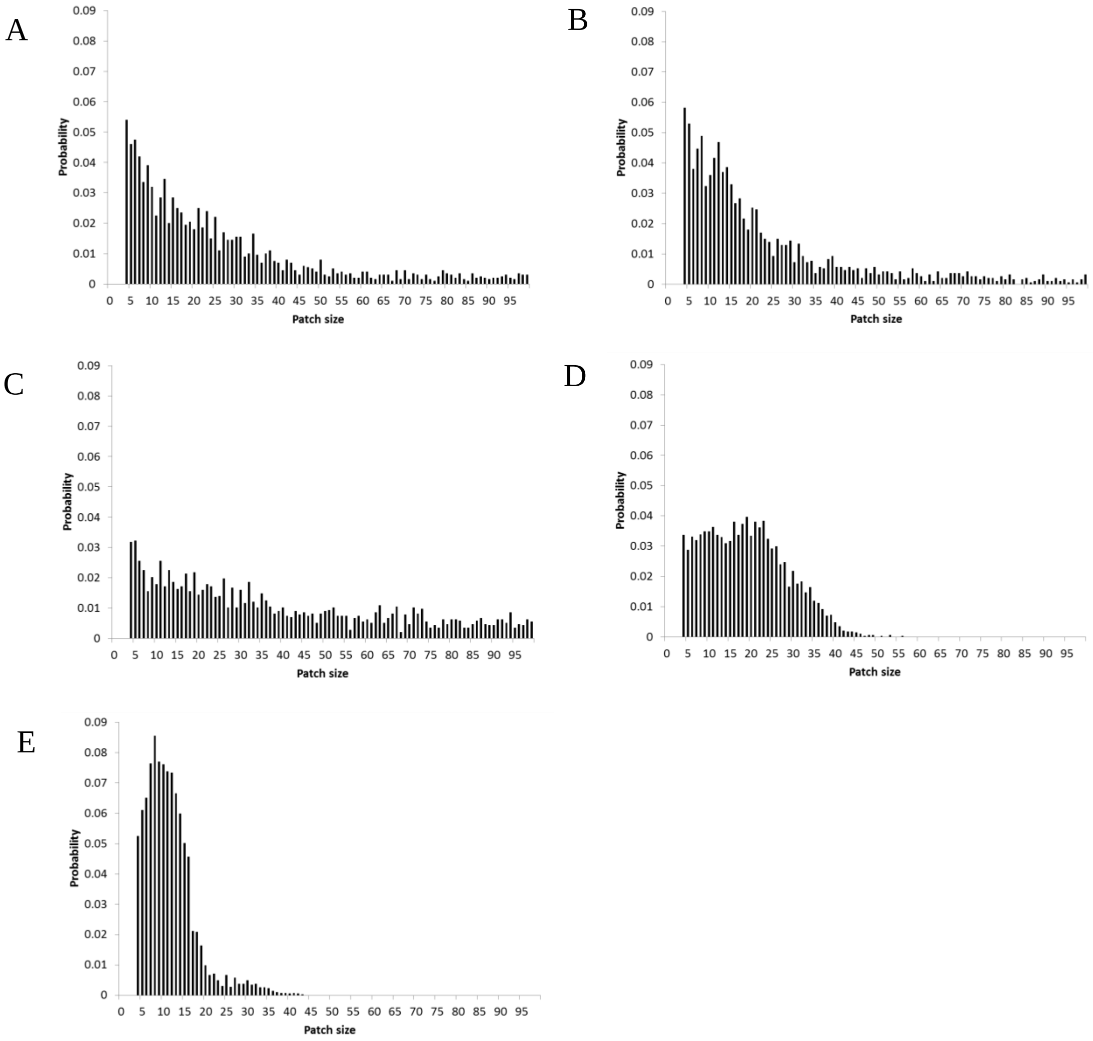


Figure 2.4. Distribution of number of patches with a specific patch size for electrostatic (A, negative, B, positive), C, hydrophobic, D, hydrogen bond donor and E, acceptor properties.

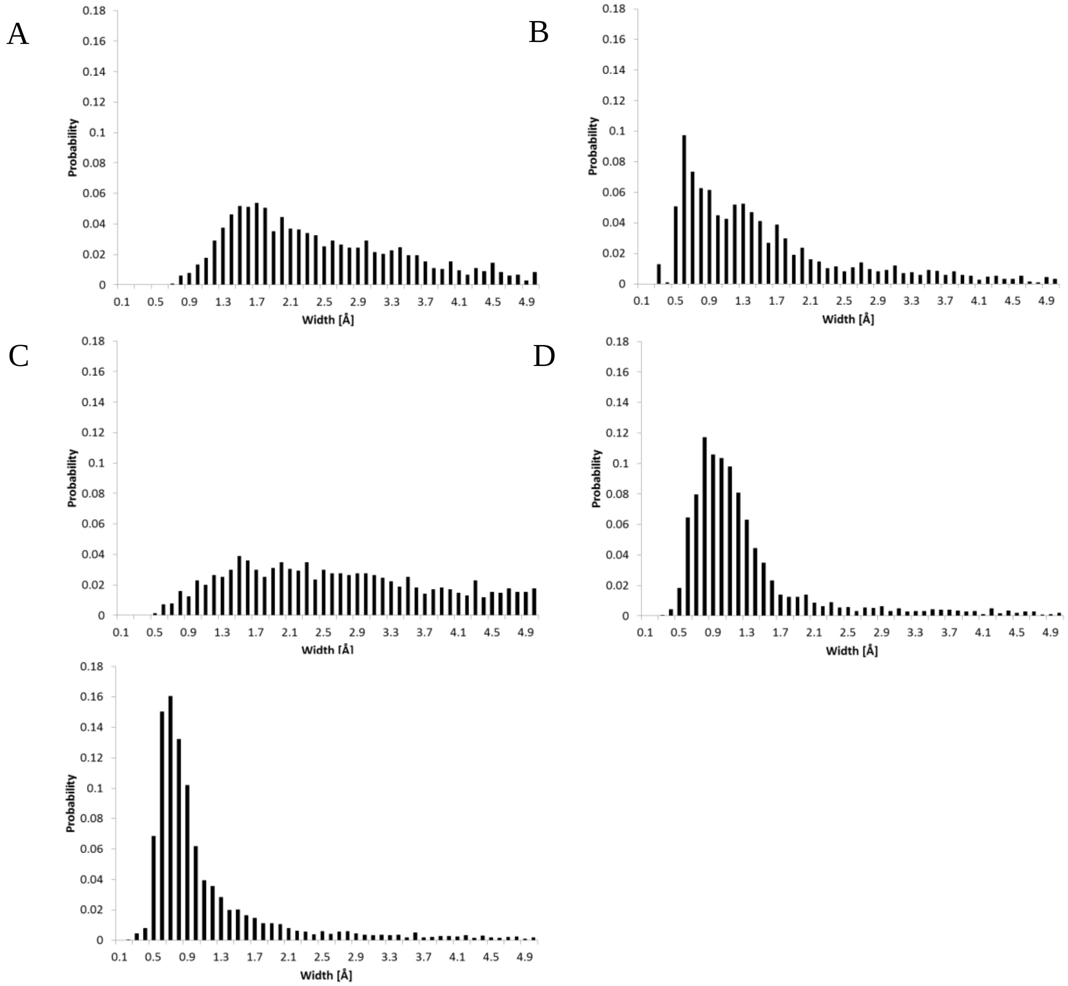


Figure 2.5: Distribution of width of Gaussian fit to patches with electrostatic (A, negative, B, positive), C, hydrophobic, D, hydrogen bond donor and E, acceptor properties.

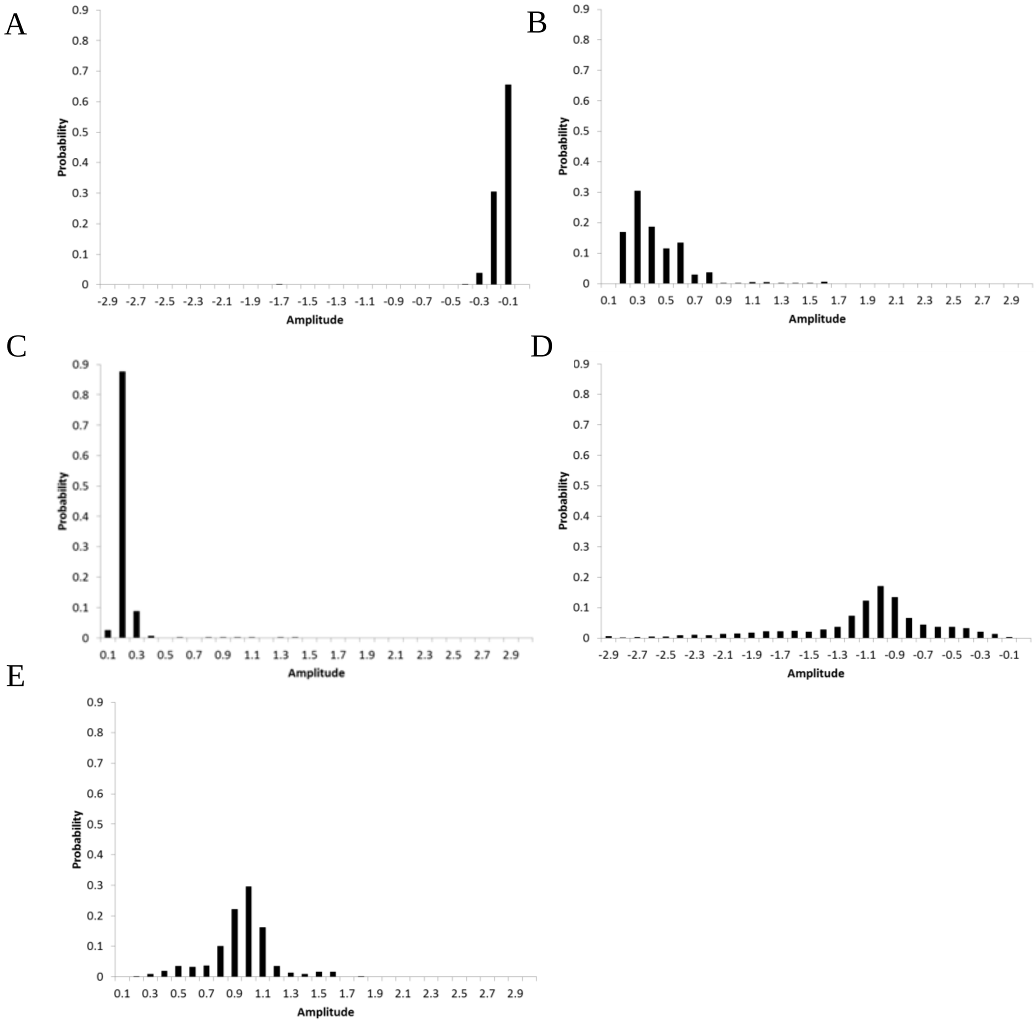


Figure 2.6: Distribution of amplitude of Gaussian fit to patches with electrostatic (A, negative, B, positive), C, hydrophobic, D, hydrogen bond donor and E, acceptor properties.

The trend for the distribution of patch sizes for the different physico-chemical properties is reproduced in the distribution of the width of the Gaussian function fit to the patches (Figure 2.5). Width has been defined as half of the sum of the lengths of the principal axes of the Gaussian function. Compared to electrostatic and hydrogen-bond properties, hydrophobic patches are on average larger in size and consequently Gaussian fits display larger widths.

Figure 2.6 displays the distribution of amplitudes of the Gaussian fits for the various properties. Interestingly, on average the negative electrostatic potential is smaller in magnitude than the positive electrostatic potential. This reflects the smaller van der Waals radius of partially positive hydrogen atoms compared to partially negative nitrogen, oxygen, sulphur or carbon atoms. Thus, surface points on the solvent accessible surface are on average closer to positive atoms than to negative atoms, which results in stronger positive electrostatic potential compared to negative potential. Most hydrogen bonding patches have amplitude of around one or minus one consistent with the maximum hydrogen bond strength of one according to equation 2.2. Hydrogen-bond acceptor patches with amplitude up to two and donor patches with amplitude up to three have been identified. Those patches represent surface regions that share multiple nearby hydrogen-bonding functional groups of different amino-acid residues that cannot be cleanly separated by the closest protein atom classification.

2.4.3 Similarity of binding sites

We also investigated how similar the identified patches and Gaussian fits were for different crystal structures of the same protein. For this study, we ran the PLSIA program on a set of structures for 4 different proteins: Estrogen receptor, CDK2, HIV protease, and RARγ. The estrogen receptor set includes both protein structures with bound agonists and antagonists. We performed a clique detection analysis of the distances between centers of the surface patches (see section 2.2.8) for all pairs of protein structures. The results were evaluated for pairs of structures of the same protein system and different protein systems (Figure 2.7).

For the RARγ and agonist-bound estrogen receptors, there is clear separation between the similarity scores (equation 2.5) for comparisons between structures of the same protein system with respect to comparisons to structures of other protein systems. Such a clear separation was not identified for the other protein systems, though for CDK2 and HIV protease the intra-protein scores are slightly higher relative to the comparisons with the structures of other protein system. For antagonist bound estrogen-receptor structures, a separation to comparisons with other protein systems is observed but not to comparisons with agonist-bound estrogen structures. This is not surprising, as the antagonists bind also to the agonist binding site but their chemical structure typically extends to a solvent-exposed moiety. As the used similarity measure (Equation 2.5) is normalized to the minimum number of patches in either of the two compared protein structures (here the agonist bound structures), the partial overlap of agonist and antagonist binding sites resulted in a comparable range of similarity scores.

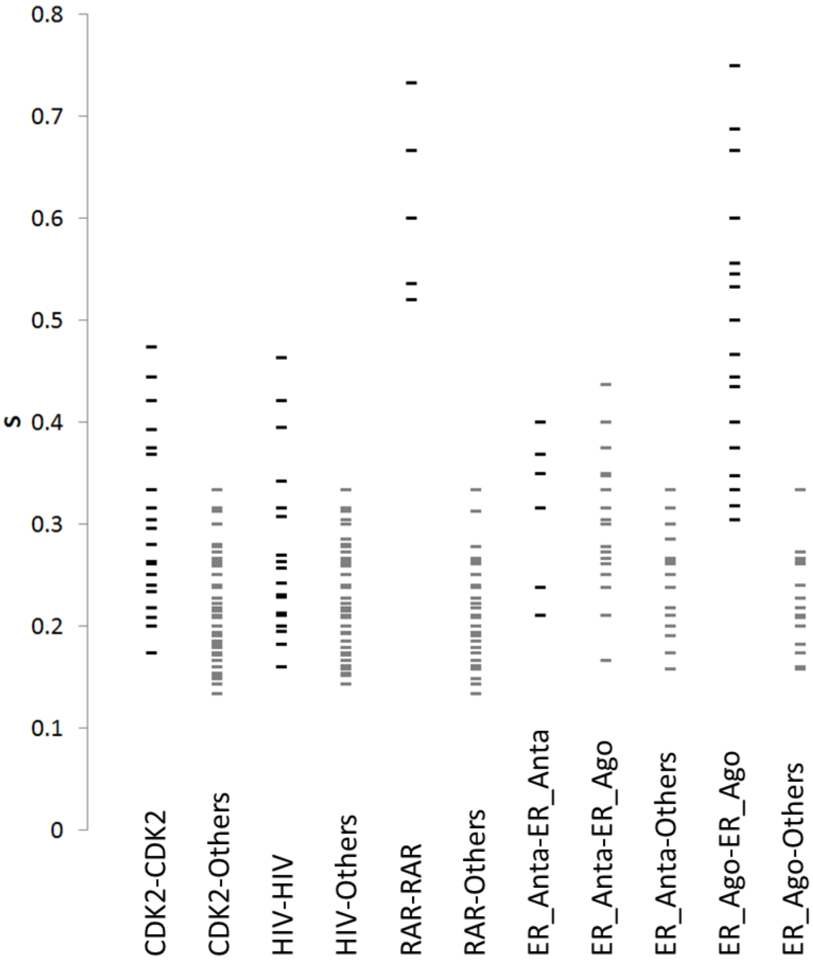


Figure 2.7: Similarity score (Equation 2.5) for all pairs of protein structures using clique detection on the centers of surface patches. Each column represents all pair-wise interactions for the indicated protein groupings, either intra or inter-protein.

Analysing each individual protein structure (Figure 2.8) reveals that in all cases the average similarity score to any protein structure from the same protein class is higher than the corresponding average score to any other protein structure. Thus, the analysis still seems to preferentially select members of the same protein class against other protein systems, even for CDK2 and HIV protease.

Visual comparisons of pairs of protein structures of the same protein system, reveals that low similarity scores are often associated with conformational changes in the binding site (Figure 2.9). The structures of the binding sites of agonist-bound estrogen receptors are relatively similar, resulting in comparable locations of the Gaussian centers and consequently large similarity scores. In contrary, the CDK2 system shows significant conformational variation in the binding pocket, which leads to dissimilar Gaussian center locations and low similarity scores.

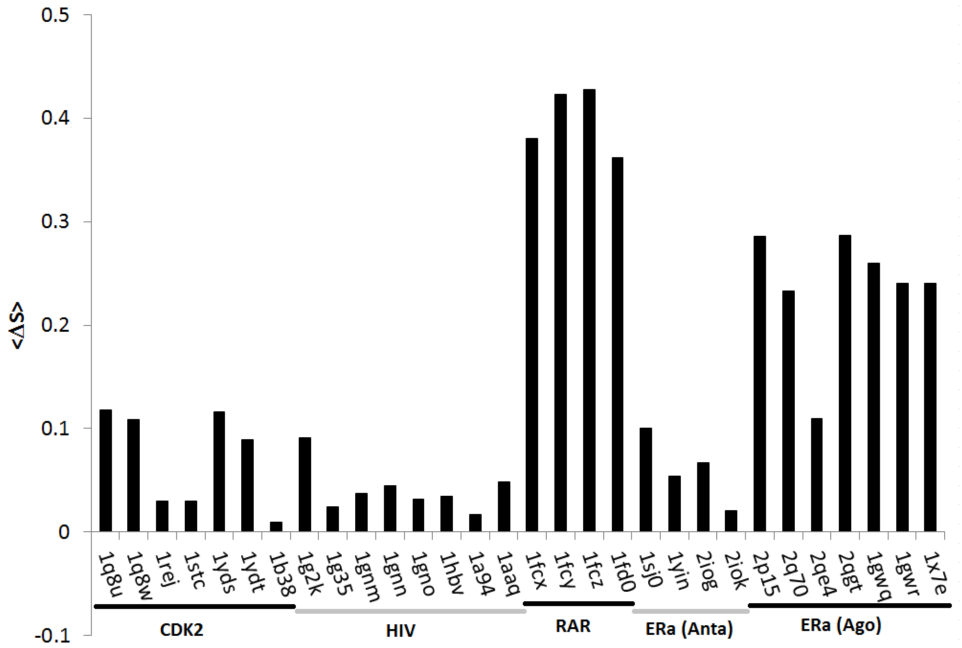


Figure 2.8: Difference in average similarity score (equation 2.5) between comparisons to protein structures within and without the same protein class for all protein structures. Positive values correspond to larger similarity among members of the same protein class. Every individual protein structure was more similar to members of its protein class (represented by a positive score) than to other protein classes. The more rigid or conserved the binding pocket, the more positive the score.

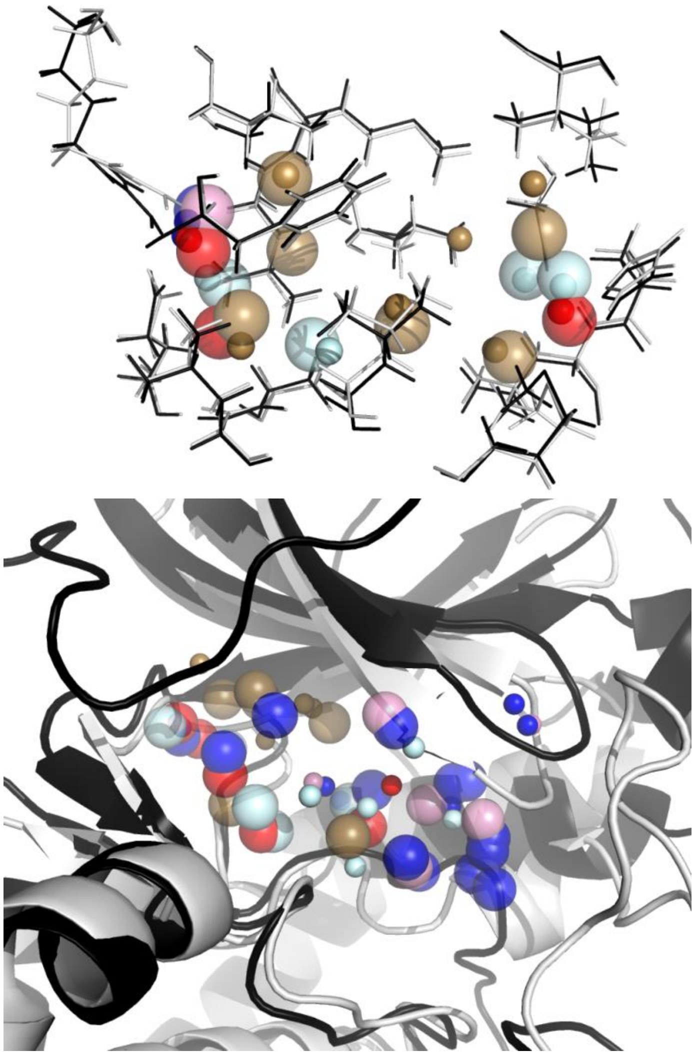


Figure 2.9: Pairwise comparison of Gaussian centers for two protein structures for two different protein systems, agonist-bound estrogen receptor (top) and CDK2 (bottom). Top: Binding site residues for the estrogen receptor structures 1gwr (black) and 1gwq (grey) are displayed as lines, the corresponding Gaussian centers are shown as small solid spheres (1gwr) and large transparent spheres (1gwq). The binding site residues don’t display significant conformational changes, resulting in similar positions of Gaussian centers. Bottom: The extended loop regions in 1rej (black) and 1b38 (grey) show significant conformational differences resulting in poor overlap between Gaussian centers of 1rej (small solid spheres) and 1b38 (large transparent spheres). Gaussian centers for electrostatic negative patches are colored cyan, positive (pink), hydrophobic (brown), donors (blue) and acceptors (red).

2.5 PLSIA Conclusions

In this chapter we studied if the physico-chemical properties of the binding site of a protein can be accurately represented by surface descriptors modelled by 2-D Gaussian functions fitted to surface patches. Properties such as electrostatic and hydrophobic properties are accurately fitted using Gaussian functions with an average relative error around 15%. Hydrogen bond properties are more localized but display larger errors around 20-25% compared to other properties. One contribution to this increased error is that hydrogen bonds are directional but the vector of directionality of a hydrogen bond donor or acceptor doesn't necessarily point in the direction of the surface normal. Adding a directionality term to the Gaussian fit function may reduce the error of fit but increases the number of fit variables and consequently the potential of overfitting.

The type and location of the Gaussian centers is consistent among different structures of the same protein system, if no significant conformational changes are observed in the binding site upon binding of different ligands. On average, only about six Gaussian function descriptors are necessary to model each physico-chemical property important for ligand binding. This demonstrates the potential to use 2-D Gaussian functions in surface-based pseudoreceptors allowing for a significant reduction of number of descriptors in the QSAR modeling process.

***Note: Portions of this chapter previously published in:

Wilson, GL.; Lill, MA. Towards a realistic representation in surface-based pseudoreceptor modelling: a PDB-wide analysis of binding pockets, *Molecular Informatics*, **2012,** *31*, 259-271

List of References

1. G. L. Wilson, M. A. Lill, *Future Med. Chem.* **2011**, *3*, 735-750.
2. Y. Tanrikulu, G. Schneider, *Nat. Rev. Drug Discov.* **2008**, *7*, 667-77
3. R. Wang, X. Fang, Y. Lu, S. Wang, *J. Med. Chem.* **2004**, *47*, 2977-80.
4. R. Wang, X. Fang, Y. Lu, C.-Y. Yang, S. Wang, *J. Med. Chem.* **2005**, *48*, 4111-9.
5. J. M. Word, S. C. Lovell, J. S. Richardson, D. C. Richardson, *J. Mol. Biol.* **1999**, *285*, 1735-47.
6. Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, et al., *J. Comp. Chem.* **2003**, *24*, 1999-2012.
7. M. L. Connolly, *J. Mol. Graphics* **1993**, *11*, 139-41.
8. S. Aragon, *J. Comp. Chem.* **2004**, *25*, 1191-205.
9. S. A. Wildman, G. M. Crippen, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.
10. M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, F. Rossi, *GNU Scientific Library Reference Manual - Third Edition*, **2009**.
11. M. Lill, A. Vedani, M. Dobler. *J. Med. Chem.* **2004**, *47*), 6174–6186.

CHAPTER 3. OPTIMIZING SURFACE REPRESENTATIONS OF BINDING SITES USING EXPERIMENTAL PROTEIN-LIGAND STRUCTURE DATA

3.1 Overview

As mentioned in Chapter 1, there are several major classes of pseudoreceptor methods including atom-based, fragment-based and residue-based methods[1,2]. One major class is surface-based methods, where the pseudoreceptor is represented as a curved 3-D surface with physicochemical properties mapped onto it representing protein properties important for protein-ligand interactions[3-6]. These surfaces are generated in a number of ways. In Receptor Surface Models (RSM), a “shape field” for each ligand is generated that represents the molecular volume[3-5]. The fields for all ligands are then combined, and an iso-level surface is generated based on the combined shape field. In RAPTOR, an iso-surface approximating the solvent-accessible surface of the aligned ligand-set is generated[6]. The occupancy of every ligand atom is mapped onto a grid according to a smooth function ranging from 1 at the atom center to 0 at its solvent accessible surface. An iso-level surface is then generated again, similar to the RSM approach. Atom-based approaches use similar methods to determine where to place the atoms of the pseudoreceptor[7-11]. For example, FLARM generates a spherical grid around the geometric center of the aligned ligands[8]. The sphere is then contracted towards the center until a grid point contacts the surface of a ligand atom. The surface is finally relaxed to allow for a cushion distance of less than one Å between ligand and pseudoreceptor. In WeP[11], the marching cube algorithm is used. The space surrounding the ligand set is divided into cubes and steric overlap of a methyl groups (2.0 Å probe radius) placed on each cube vertex with the ligands is tested. The interacting cube vertices are then used for generating a triangulated surface representing the pseudoreceptor.

Whereas those different schemes aim to empirically reproduce the surface of the binding site using ligand information only, to the best of our knowledge, no systematic investigation was performed to validate which surface generation process most accurately reproduces the surface of the real binding site of experimentally determined protein-ligand structures.

In general, experimental information about the protein structure is not used to generate a pseudoreceptor. In this study, however, we will use experimental data to optimize the pseudoreceptor method to accurately represent the binding pocket for any given protein. We studied a number of protein-ligand crystal structures for three different protein systems and investigated whether the molecular surface of the protein structure can be reproduced with iso-surfaces generated from the corresponding co-crystallized ligands. Throughout our analysis, we have identified a set of parameters that reliably reproduces the surface of the binding pocket for all studied protein systems.

3.2 Methods

3.2.1. Protein Surface Generation

For the protein-surface analysis 35 protein-ligand structures from three protein systems were selected from the Protein Data Bank (PDB) (www.rcsb.org)[12]: 20 cyclin-dependent kinase 2 (CDK2) structures, 7 estrogen receptor α (ER) structures, and 8 HIV protease (HIV-PR) structures. (Table A1) The crystal structures within each protein system were aligned to each other using PyMOL. The molecular surface of each protein binding pocket was then identified using MSRoll in conjunction with our previously described refinement algorithm[13]. (Figure 3.1a) For ER, surface points with a maximum distance of 4 Å to any ligand atom was used; for CDK2 and HIV-PR this cut-off was set to 5 Å. The lower cut-off for ER was necessary to prevent the inclusion of surface points in our analysis that result from cavities other than the binding pocket.

3.2.2. Occupancy Calculation

The crystal structure ligands were extracted from the aligned PDB structures and grouped into five categories: ER, HIV-PR, CDK-20, CDK-10, and CDK-5. The ER, HIV-PR, and CDK-20 groups contain all ligands for each respective protein system, while CDK-10 and CDK-5 are randomly chosen subsets of the CDK-20 group made up of ten and five ligands respectively. A grid around the aligned ligand molecules was constructed using the minimum and maximum x-, y- and z-values, (xmin, ymin, zmin) and (xmax, ymax, zmax), respectively of any atom of the ligand set plus an additional 10 Å cushion in each positive and negative direction. Grid points were placed starting from (xmin, ymin, zmin) using a grid-spacing of 1 Å.

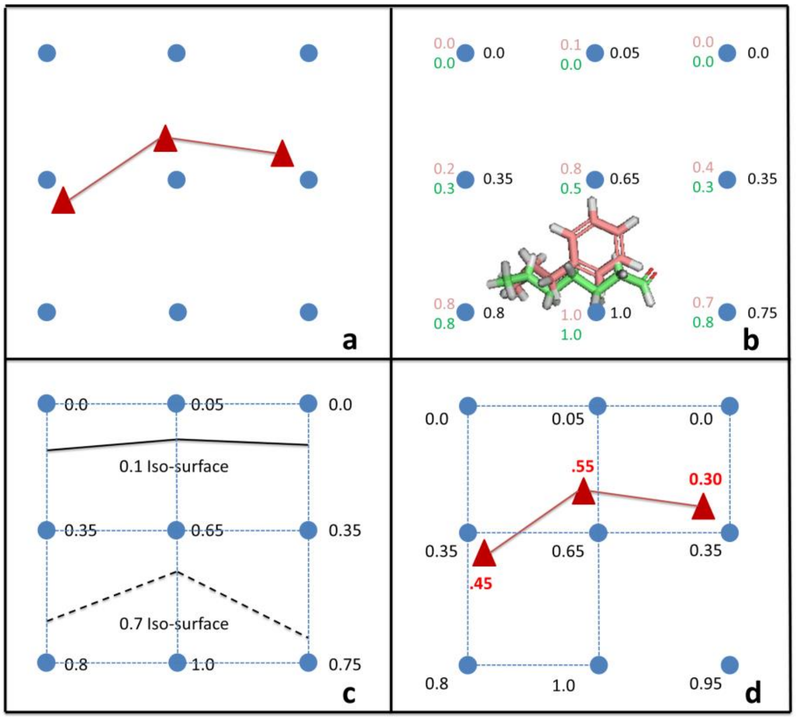


Figure 3.1: Schematic of algorithm: (a) Generation of triangulated protein surface. Red triangles and lines are vertices and edges of triangulated MSRoll surface. Blue circles represent grid points generated in subsequent steps. (b) Occupancy of ligand molecules mapped to grid. Occupancy of ligand 1 (pink) and ligand 2 (green) are calculated on grid points. Occupancy is averaged across all ligands to produce final value (black). (c) Iso-surfaces of ligand occupancy are generated. Occupancy values are interpolated between grid points to match a target iso-level. These vertices are then used to generate an iso-surface shell of ligand occupancy. The solid black line represents the 0.1 occupancy iso-surface, the dashed line the 0.7 occupancy iso-surface. (d) Interpolation of protein surface occupancy. Protein surface points (red triangles) are placed inside the grid generated from their corresponding crystal ligands (blue circles). The ligand occupancy values of these grid points are then interpolated to generate occupancy values for each protein surface point

The steric occupancy of the ligand atoms were then mapped onto the grid. These occupancy values represented the shared volume there would be between a sphere placed on the grid point with the molecular (van der Waal's) volume of a ligand atom, with a value of one representing full overlap, and zero indicating no overlap. Occupancy was calculated with the same function as used in the RAPTOR QSAR package[6]:

$$O = 1 - \left(\frac{4}{9}\right) * \left(\frac{d}{M}\right)^6 + \left(\frac{17}{9}\right) * \left(\frac{d}{M}\right)^4 - \left(\frac{22}{9}\right) * \left(\frac{d}{M}\right)^2 \qquad d < M \qquad (3.1)$$

$$O = 0 \qquad d \geq M$$

$$M = r_{vdW} + c$$

Where O is the occupancy, d is the distance between ligand atom and grid point, and M is the maximum radius. The maximum radius for a ligand atom is defined as the van der Waal's radius r_{vdW} of that atom plus a constant value *c*. In this work, four different constants were used: *c* = 1.4 Å, 2.0 Å, 2.5 Å, and 3.0 Å. (Figure 3.2) The 1.4 Å constant reflects the solvent accessible surface (SAS) of a ligand, as the occupancy function reaches zero at the SAS radius using this constant.

For every ligand molecule, the occupancy for each atom was computed on all grid points within 4 Å plus the van der Waals radius of the atom. (Figure 3.1b) The highest atom occupancy for every given grid point was stored as the final occupancy value for the ligand molecule. In the cases of the 1H00 structure for CDK2, there were two ligand conformations present, so the occupancy values for the grids of the two conformations were averaged to give a final occupancy for that ligand. The occupancies for all ligand molecules within a set were then averaged to produce a final occupancy for each grid point representing the full ligand group.

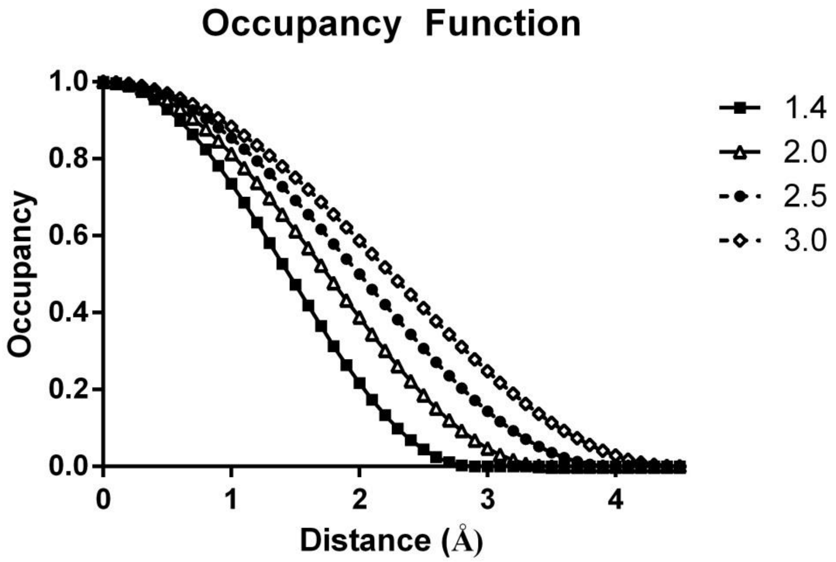


Figure 3.2: Occupancy as function of distance with varied values of c. Function begins at a value of one at zero distance and decays to zero at a distance equal to the sum of the van der Waals radius of an atom plus a constant c. In this graph, the van der Waals radius is set to 1.5 Å for all value of c

3.3. Analysis

The occupancy grid was used for the final two steps of the analysis: comparison with the protein surfaces and construction of iso-surface shells. For a given iso-level, the iso-surface was constructed using the marching cubes algorithm[14]. First, the occupancy grid was searched in a systematic manner, starting with the origin point of the grid (xmin, ymin, zmin). The seven vertices surrounding this point in the positive x, y, and z directions were identified and used to generate a cube. If all eight vertices of the cube have occupancy values higher or lower than the target iso-level, the cube was discarded and the next cube was searched. This process continued until the full grid had been searched and all occupancy values mapped.

When a cube had at least one vertex with occupancy higher than the target iso-level and at least one vertex with occupancy lower than the target iso-level, it was identified as a "surface cube" at a given iso-surface level. The marching cubes algorithm then determines the intersection of the target iso-surface with the cube. This was done by interpolating where the edges of the cube intersect with the iso-surface. These intersections are then used to determine one or more surface triangles representing the target iso-surface. These triangles were then stored, and once all cubes had been searched, combined into a single triangulated iso-surface shell of the ligand occupancy. This is illustrated in Figure 3.1c. The grid in the figure has been divided into four squares representing cubes of the 3-D grid, and two example iso-surfaces. The 0.1 iso-level value iso-surface passes through the top two squares, as they have lower vertices greater than 0.1 and upper vertices less than 0.1, just as the 0.7 iso-surface passes through the bottom squares for the same reason.

An iso-surface with target value between 0.35 and 0.65 would pass through all four squares, as each has at least one vertex above and one vertex below that value.

The final process of the algorithm was to compare the known protein surfaces generated from the crystal structures with the occupancy grid generated from the co-crystalized ligands of those structures. To achieve this, the surface points, generated from MSROLL, for each protein structure of a set were placed into the generated occupancy grid of the ligands. The eight grid vertices of the occupancy grid surrounding every surface point were identified, and tri-linear interpolation was used to determine the occupancy value at the coordinates of the surface point. (Figure 3.1d) This value represents which iso-surface shell of the ligand would pass through each given protein surface point. This process was repeated for every protein structure of a given set.

Histograms were generated to determine how well an iso-surface a given iso-level is able to reproduce the experimental protein surface. The histograms measure the percent of protein surface points that are spatially congruent with the iso-surface at a given iso-level of occupancy displayed on the x-axis of the graphs. When discussing the graphs, coverage percentage refers to the percent of protein surface points that would be contained within an iso-surface shell of a given iso-level. Cumulative occupancy graphs were generated to display the coverage percentage.

3.4 Results

The primary focus of this study is to derive an optimized algorithm to generate pseudoreceptor surfaces that closely mimic the experimental binding pocket surface of protein structures using the co-crystallized ligands. To achieve this end, we performed both a quantitative as well as a qualitative analysis of the results of our algorithm, generating the following results for the individual protein systems.

3.4.1 Estrogen Receptor

Figure 3.3 shows the histogram results for the surface comparison analysis using the four different constants c for the ER protein set. The results for the 2, 2.5, and 3 Å c-values show similar profiles. The higher c-value histograms show shifts towards higher average occupancy values. This is due to the fact that once a surface point lies within the maximum distance of the occupancy function an increase in the c-value will simply result in an occupancy value closer to one. There are slight differences between these three constants: instead of the single maximum in the histogram using c = 2.0 Å, for c = 3.0 Å there is a small additional maximum at lower occupancies. This feature starts to appear in the c = 2.5 Å statistics. This is most likely caused by ligand variation. One ER ligand is significantly different from the rest, with a group occupying a unique region of space. This difference causes a large variation in the protein surfaces, and causes the double maxima, as there are two distinct surface profiles. For c = 1.4 Å, a significantly different profile is observed with a small number of points located outside the maximum distance: approximately 0.15%. This means that these points are located more than the c-value plus van der Waal's radius away from any ligand atom. The width of the histogram peak is also very compressed, with 25% of points located between occupancies of 0 and 0.15 with the maximum located in the bin between 0.05 and 0.10 occupancy.

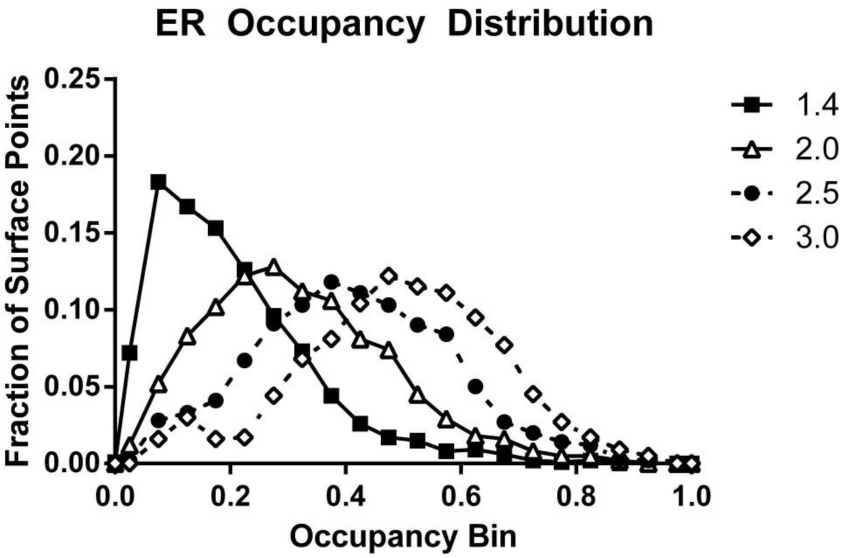


Figure 3.3: ER occupancy distribution graph with varying values of c in occupancy calculation. Occupancy values are binned in .05 width increments, starting with 0.025 as a bin center representing the 0-0.05 bin and increasing to 0.975. Bins are inclusive on the upper limit, exclusive at the lower limit. An additional zero bin is added which includes the fraction of protein surface points with no mapped occupancy

Figure 3.4 shows the cumulative histogram for the ER system. Of particular interest are the occupancy bounds for the majority of the surface points (i.e. finding the ligand iso-surface shells that would surround the majority of surface points. For c = 1.4 Å, 93% of protein surface points have interpolated occupancies higher than 0.05, and 74.4% higher than 0.10. On the other tail, only 1% of points have occupancies higher than 0.65, 6% higher than 0.45, and 13% higher than 0.35. Together, this means 80% of surface points are located between 0.05 and 0.35 occupancy, 90% between 0.05 and 0.55 and 99% between 0 and 0.65. For the other values of c, the occupancy iso-levels with the same percent coverage increase with c-value. This occurs through all coverage percentages, increasing with greater iso-level values, leading to an increase in the difference of iso-level values between any two coverage percentages. For example, 94% of points have higher occupancies than 0.1, 0.15, and 0.2 for the 2.0, 2.5, and 3.0 Å constant values, respectively. The 5% coverage iso-levels are for these runs: 0.6, 0.7 and 0.75. So the total separation for the iso-surface shells containing between them 90% of protein surface points increases from 0.5 occupancy difference for 1.4 Å to 0.55 for the 3.0 Å run. The 99% separations are much larger: 0.65, 0.8, 0.85, and 0.9 for the 1.4 Å, 2.0 Å, 2.5 Å, and 3.0 Å c-values.

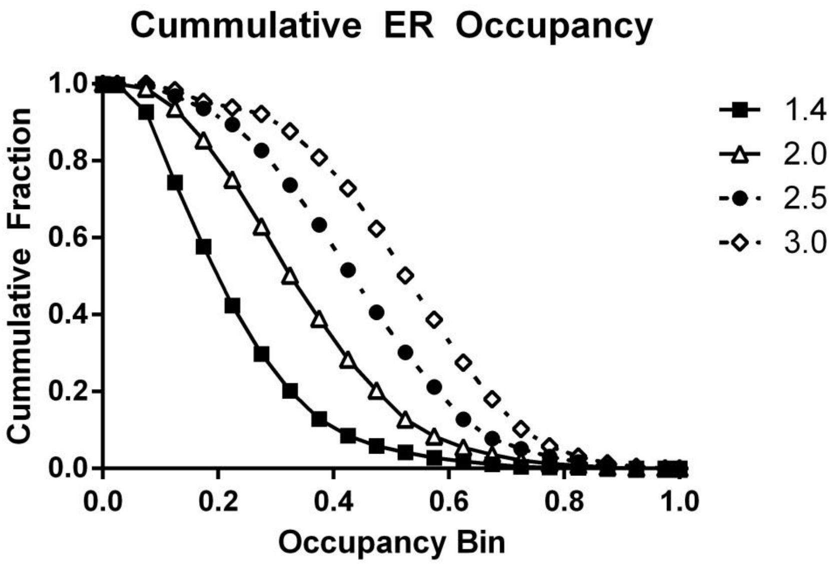


Figure 3.4: ER accumulation graph with varying c-values in occupancy function. Y-axis is percent of total protein surface points with iso-level of less than or equal to the x-axis bin. Bins are determined in the same manner as the corresponding distribution graph

3.4.2 HIV-PR

The results for HIV-PR are shown in Figure 3.5. The results for this system follow similar trends to the ER system. The distribution profile for c = 1.4 Å is significantly different from the other three c-values. New profile features start developing when c = 2.0 Å, and these features develop fully in the 2.5 Å and 3.0 Å runs, which show very similar profiles. A major difference between the results for HIV-PR as compared to ER is that it is not until c = 3.0 Å that all surface points have a higher than 0 occupancy. This is partly expected, as the distance cut-off for protein surface points is 1 Å larger for the HIV-PR and CDK systems. While the percentage of uncovered surface points is relatively high, 5%, for c= 1.4 Å, it decreases to 1% and then to 0.08% for c = 2.0 Å and c = 2.5 Å. The bimodal motif that was partly evident in the ER set is much more pronounced in the HIV-PR set when c= 2.5 and 3.0 Å, with a swift increase at low occupancies that plateaus for a significant range of occupancies, and then increases to a maximum followed by a decrease, as seen in Figure 3.5.

The accumulation results for the HIV-PR set are shown in Figure 3.6 and Table A5. For c= 1.4 Å a significant fraction of points have low occupancies. 20% of protein surface points have occupancy less than 0.05. 99% of points have occupancy lower than 0.55, 95% lower than 0.45, and 90% lower than 0.35. For c = 2.0 Å, 99% of points have occupancy > 0.0, and 90% have occupancy higher than 0.05.

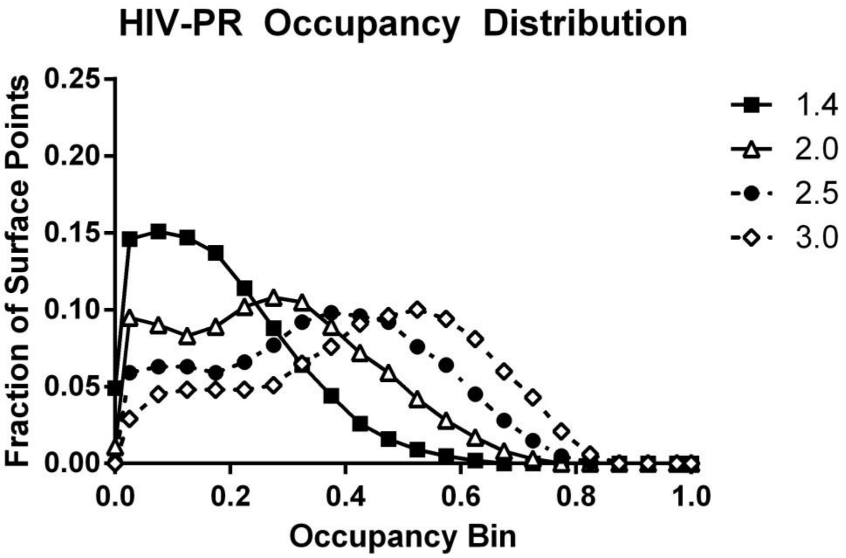


Figure 3.5: HIV-PR occupancy distribution graph with varying values of c in occupancy calculation. Occupancy values are binned in .05 width increments, starting with 0.025 as a bin center representing the 0-0.05 bin and increasing to 0.975. Bins are inclusive on the upper limit, exclusive at the lower limit. An additional zero bin is added which includes the fraction of protein surface points with no mapped occupancy

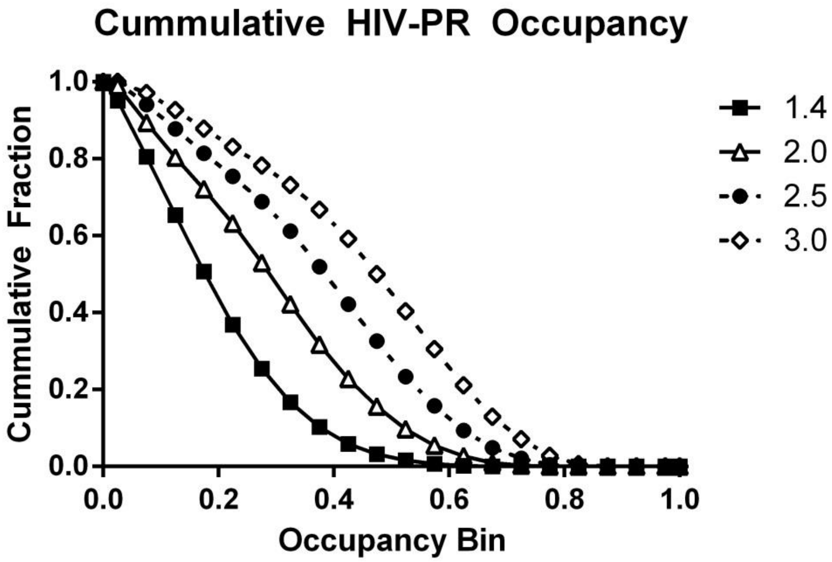


Figure 3.6: HIV-PR accumulation graph with varying c-values in occupancy function. Y-axis is percent of total protein surface points with iso-level of less than or equal to the x-axis bin. Bins are determined in the same manner as the corresponding distribution graph

3.4.3 CDK

In this work, three different sets of CDK2 ligands were used to produce occupancy data: CDK-20, CDK-10, and CDK-5. CDK-10 is a subset of CDK-20 and CDK-5 is a subset of CDK-10. In the analysis stage, however, all twenty protein structures were used for all three ligands sets. This was done to determine how well the iso-surface shells from a smaller ligand set correspond to the larger ensemble of protein structure. Figure 3.7 and Tables A6-A11 show the results for the individual sets. (Occupancy accumulation graphs for CDK are not shown) Whereas the total range of occupancy values of the protein surface points is similar in all sets compared to the HIV-PR and ER system, there is a notable shift in the distribution of occupancy values towards higher median occupancies for all c-values higher than 1.4 Å. This shift in average occupancy increases with the size of CDK set, with the CDK-5 set having the lowest median occupancy and CDK-20 the highest.

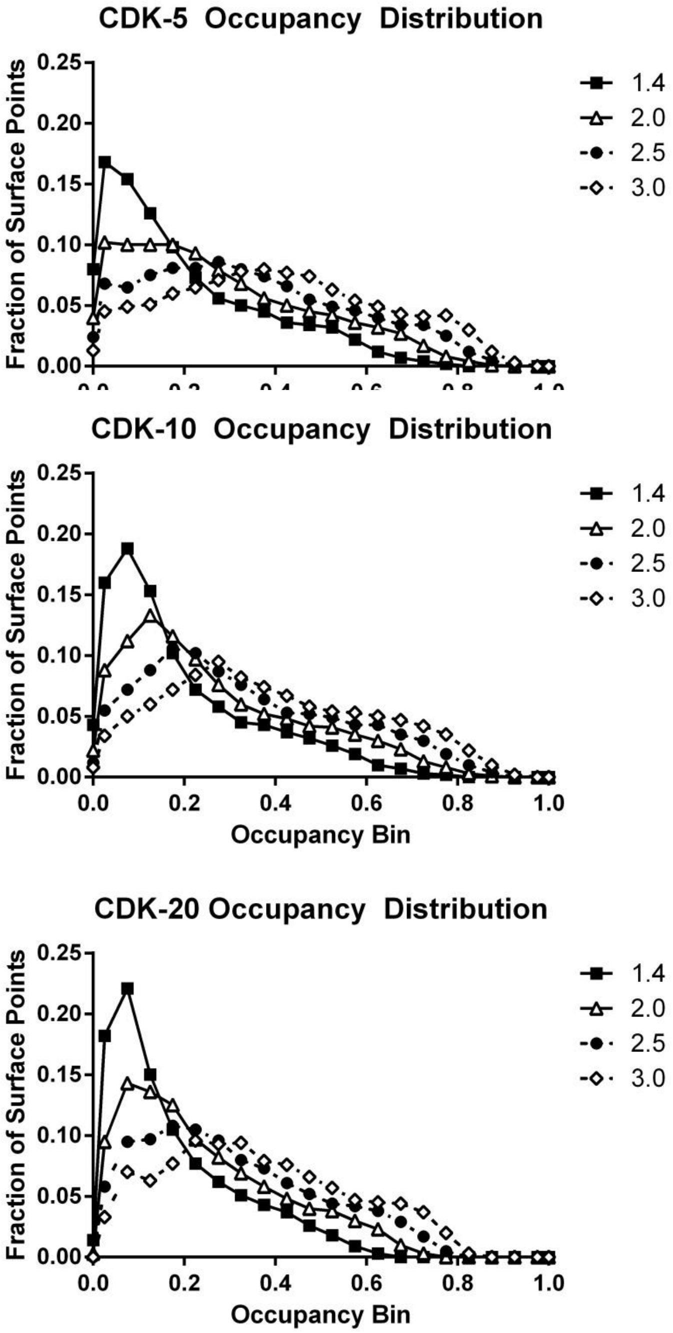


Figure 3.7: CDK occupancy distribution graphs with varying values of c in occupancy calculation for all three CDK systems. Occupancy values are binned in .05 width increments, starting with 0.025 as a bin center representing the 0-0.05 bin and increasing to 0.975. Bins are inclusive on the upper limit, exclusive at the lower limit. An additional zero bin is added which includes the fraction of protein surface points with no mapped occupancy. (A) CDK-5 set (B) CDK-10 set, and (C) CDK-20 set

Looking at the four different c-values across all three sets, there are a number of similar trends. First, the larger the ligand set, the higher the maximum. Second, the smaller the ligand sets the more surface points are located at the extreme occupancies: more points with very low occupancy and more points with very high occupancy. For the high occupancy points, this is due to the averaging occurring in the occupancy calculation. As more ligands are included, with spatial and chemical diversity, the high occupancy iso-surface shells decrease in volume. In addition, at very high iso-levels, approximately 0.9, the algorithm fails to build a full continuous shell. This is due to the averaging process across diverse ligands leading to low maximum occupancies. For example, if a set was comprised of two non-overlapping ligands, the maximum possible occupancy would be 0.5 due to the averaging process. The reason for the higher number of very low occupancy points is the inverse of this process. As long as a surface point falls within the cut-off distance of a single ligand atom, it receives a non-zero occupancy value, even if the averaging process makes it very small. This process shifts points to lower occupancies overall, which results in the increasing maxima, which are located at low occupancy values in all three CDK sets.

3.4.4 Iso-surface Shells

While looking at the previous histograms provides information on how many protein surfaces points are covered by a given iso-surface shell, it does not provide the whole story on the quality of the fit. For example, a 50 Å sphere centered on a ligand set would likely provide 100% coverage, while being of poor use in pseudoreceptor modeling. To address this issue, a number of iso-surface shells of the ligands were generated for visual inspection. Similar to the previous section, we wanted to investigate the effect of changing the distance constant c, the number of ligands used in shell generation, and target iso-level on the overall shape and size of the iso-surface. When referring to these figures, exterior refers to the space that is located outside of the protein surface when viewed from the ligand center; interior refers to the space inside the protein surface, respectively. The transition region is the 3-D space where the protein surface points are located. (Figure 3.8)

Figure 3.9 shows the 0.05 iso-surface shell of the HIV-PR system at the four different c-values. The overall shape of the shells is similar for each c-value, the primary change being the spatial extension of the shell. Figure 3.13 also demonstrates how closely the shells match the protein surface. Using c = 1.4 Å, the majority of the shell is located slightly to the exterior of the protein surface points, but is in the transition region in some portions. For c = 3.0 Å, even though the iso-surface shell is located almost exclusively to the exterior, the shell is overall significantly larger than the protein surface. The intermediate c-value iso-surface shells fall between these two extremes, with decreasing portions of the shell in the transition region, but with other regions located exterior to the protein surface.

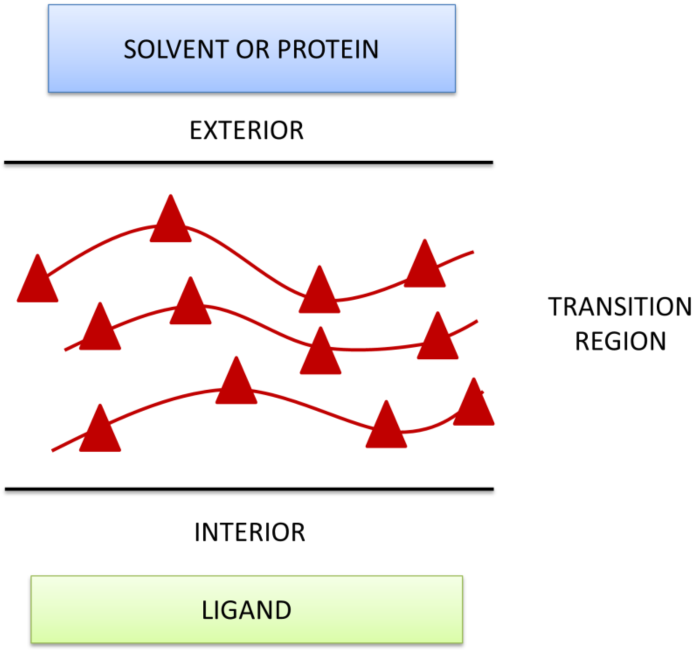


Figure 3.8: Definition of terms for discussion of iso-surface shells. Red triangles and lines represent individual protein surfaces. Black lines divide space into three regions: exterior, transition region, and interior. Interior refers to the region of space that is enclosed by all protein surfaces, corresponding to the intersection of all protein binding pockets. The transition region refers to the 3-D space where the varying protein surfaces are located. The exterior is refers to the region of space that would be filled by the protein or bulk solvent that surrounds the binding pocket

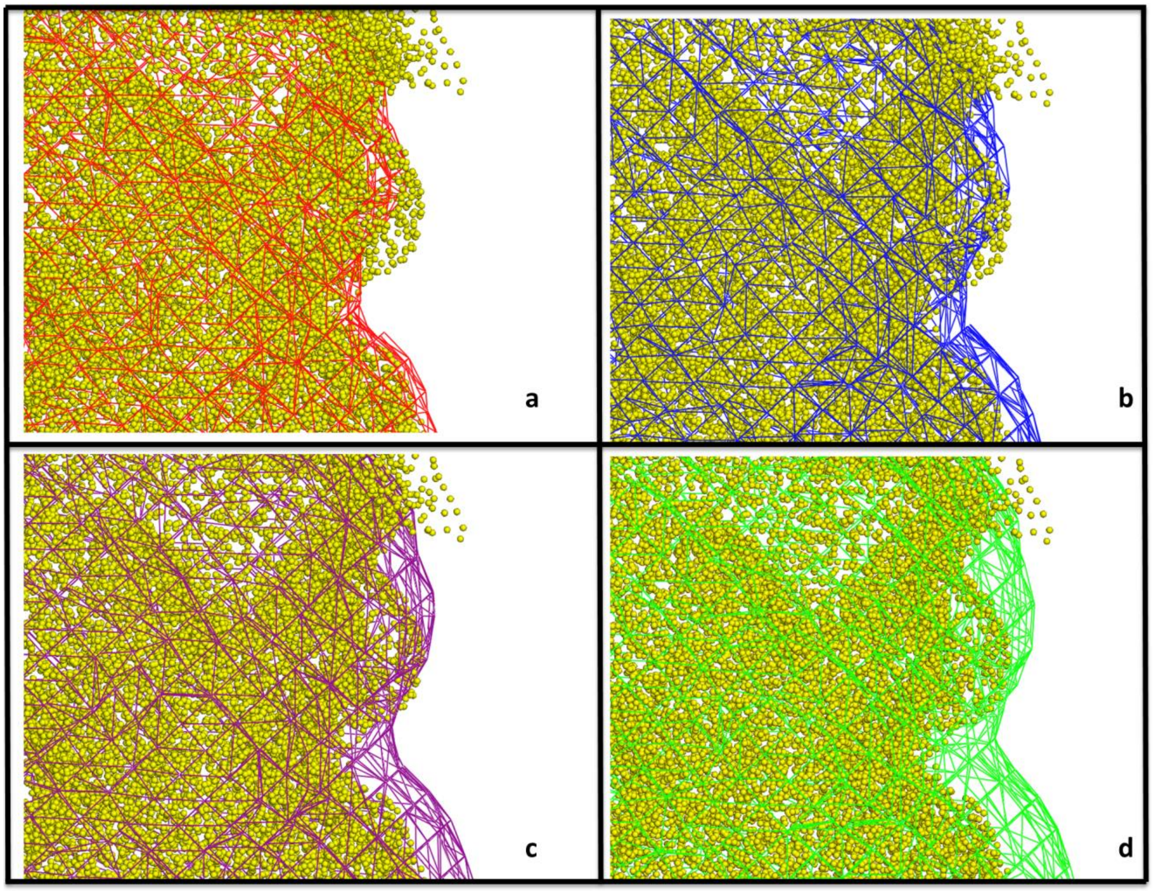


Figure 3.9: 0.05 Iso-level shells for HIV-PR with varying c-values: (a) 1.4, (b) 2.0, (c) 2.5, (d) 3.0. The iso-surfaces are generated using an iso-level of 0.05. Increasing c-value results in an expansion of the shells. All sub-figures focus on the same region of the HIV-PR binding pocket showing the larger shells including previously excluded protein surface points in one portion while simultaneously diverging from the protein surfaces in another

The effect of changing the number of ligands in the CDK systems is shown in Figure 3.10. The shells shown represent the 0.05 iso-surface with c = 2.5 Å. Overall, the three iso-surface shells are very similar. No shell is consistently larger than any of the others, though there are regions where each shell is largest. However, as the size of the ligand set increases, the curvature of the iso-surface shells becomes more refined, creating a slightly more complex surface. This consistency is desirable, as it indicates that a small ligand set can generate a pseudoreceptor that could be applicable to a larger set of ligands as long as they cover similar space in the binding pocket. This is the case for the CDK system we studied where the ligands of the CDK-5 set cover roughly the same 3-D space when aligned as the full CDK-20 set.

Figure 3.11 shows the iso-surfaces of the ER system at varying iso-levels with c = 3.0 Å. The 0.05 shell encompasses nearly all of the protein surface points and follows the contours of the protein surface. As the iso-level value increases, the encompassed volume of the shells decreases. This decrease is most pronounced in the region indicated by the red arrow in Figure 3.11b. This region is occupied by a single protein-ligand complex (2P15), and receives a very low average occupancy value, and the 0.25 iso-surface shell does not contain this region. The 0.5 iso-surface shell is mainly located in the transition region of the protein surface points. At high iso-levels, such as 0.75, the shell is almost completely to the interior of the protein surfaces.

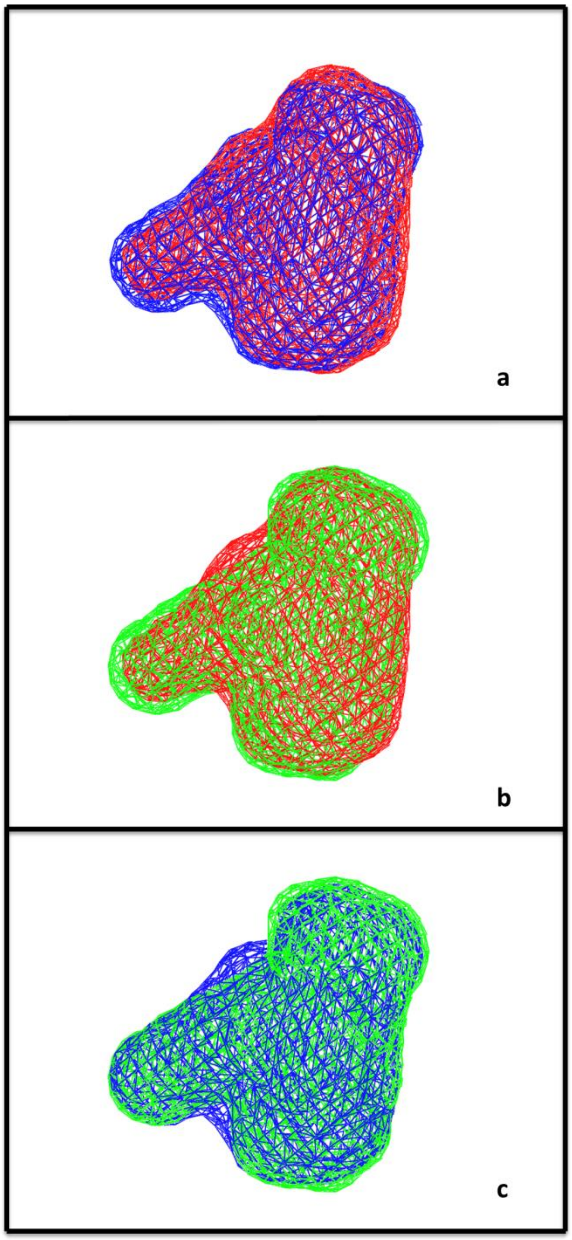


Figure 3.10: 0.05 Iso-level shells of CDK sets with c-value of 2.5 and iso-level of 0.05. (a) Paired shells for CDK-5 (red) and CDK-10 (blue), (b) CDK-5 and CDK-20 (green), (c) CDK-10 and CDK -20. While slight variations exist between all three shells, the overall shapes of the shells are very similar between the three CDK sets

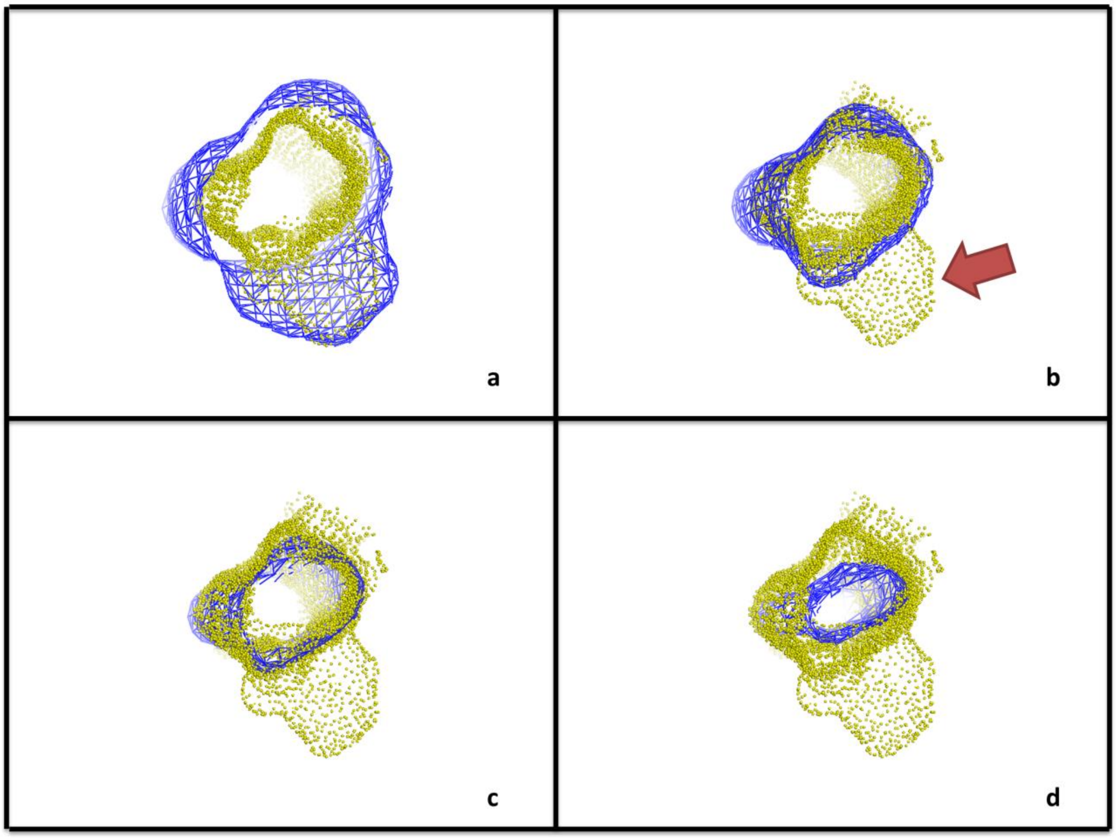


Figure 3.11: Iso-surface shells of estrogen with c=3.0. Iso-levels are (a) 0.05, (b) 0.25, (c) 0.5, (d) O.75. The size and shape of the iso-surface shell varies significantly with change in iso-level. Most notable is the change between the 0.05 shell and the 0.25 shell in the region indicated by the red arrow in sub-figure B. The protein surface points in this region come from a single protein structure, giving the region a very low average occupancy even where the ligand for that structure is located. This causes a dramatic difference between the 0.05 shell, which includes the full region, and the 0.25 shell which does not include the space

3.5 Conclusion

As previously mentioned, the focus of our research was to determine a means of producing a pseudoreceptor iso-surface that corresponds to the protein-ligand interaction surfaces present in PDB crystal structures. To achieve this goal, we first investigated how well the composite solvent-accessible surface (SAS) of a ligand set reproduces the protein surfaces, as the SAS is used in a number of pseudoreceptor generation methods. We approximate this surface by setting c=1.4 Å. At this value, the occupancy function decays to zero when a grid point is 1.4 Å plus van der Waals radius away from a given ligand atom. Therefore, any grid point with occupancy greater than zero would be within the typical SAS, while grid points with zero occupancy are outside the SAS. The SAS covers the majority of protein surface points. However, with the exception of the ER system, there remained a small portion of surface points not contained within the SAS, with a minimum coverage of 92% for the CDK-5 set. In addition, from visual inspection, the SAS iso-surface shells with low (<0.05) iso-levels, are located in close proximity to the protein surface. Conversely, increasing the c-value to 3.0 Å ensures nearly complete (>99%) coverage, but the iso-surface shells included portions in 3-D space that would overlap with the protein.

In order to create an accurate pseudoreceptor surface model, we need to find a balance between achieving the maximum possible coverage of protein surface points and smoothly approximating the protein surfaces without significant overlap with the protein. Also, from visual inspection, coverage percentage may be misleading in certain cases. When a protein binding pocket is solvent exposed, it is possible for surface points that are within the distance cut-off to be in fact outside the binding site. (Figure 3.12) Due to the opposing geometries of the protein binding pocket which are concave and the generated iso-surface which is convex, slightly lower coverage percentages are observed. This underestimation must be considered in evaluating the results of the algorithm. From our results, we would recommend a c-value of less than or equal to 2.0 Å, as higher c values produce shells that are significantly larger than the protein surfaces at low iso-levels. For these c-values, we recommend an iso-level target of <0.05, as these parameters result in iso-surface shells that cover approximately 95% of protein surface points, while smoothly approximating the protein surface in most regions.

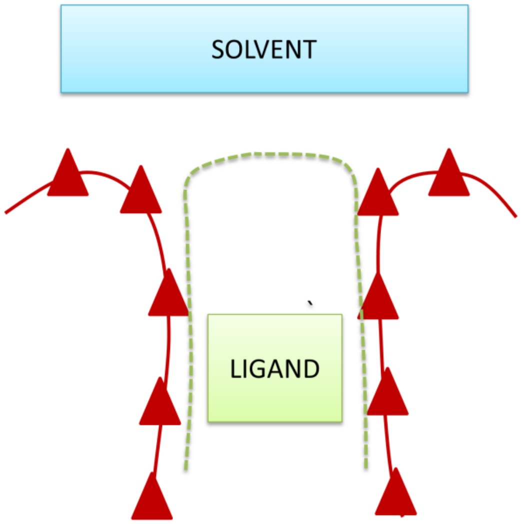


Figure 3.12: Scheme for coverage percentage of solvent exposed ligand binding pockets. Red triangles and line represent the protein surface of a single protein-ligand crystal structure. Due to curvature where the pocket is solvent exposed, certain protein surface are included within the distance cut-off of the algorithm that do not fall within the convex ligand-iso-surface shell (dashed line)

While the previous parameters for c and iso-level are useful in determining an iso-surface shell that encompasses the full combined protein surface, this shell fails to address a number of issues. First, ligand diversity can vastly increase the size of an iso-surface shell constructed from a low iso-level target, as seen in the ER system, where there is a vast difference between the 0.05 shell and the 0.25 shell due to a single ligand having a pose which occupies a different region of the binding pocket compared to all other ligands. In addition, the surfaces of the individual proteins vary significantly due to protein flexibility, resulting in a wide range of mapped occupancies (>0.5 iso-level difference range for 95% of protein surface points). With respect to pseudoreceptor modeling, this means while the low iso-level iso-surface shell represents the outermost surface to all ligands, it does not fully replicate the surface with which an "averaged" ligand would interact, especially if the ligand set is diverse. This averaged surface would be represented by an iso-surface shell closer to interior of the protein surface points, inside the transition region. As seen in Figure 3.11, in the ER system the iso-level that corresponds to this region is in the range of 0.5-0.6.

It is also important to note, that while the suggested parameters represent general starting points, they will not be ideal for all protein ligand systems. In just the three systems considered in this study, there are significant differences in the occupancy profiles of the protein surface points. The flexibility of the protein and the diversity of the ligand set play important roles in determining the ideal parameter set. A rigid system would lead to a lower ideal c-value, as seen in the ER system, where even with c=1.4, less than 0.1% of protein surface points do not have an occupancy value. Increasing ligand diversity leads adjusting the iso-level targets of the interior and exterior shells. For example, with the

CDK-20 set, it is impossible to create an iso-surface shell with iso-level greater than around 0.7, as the surface becomes discontinuous due to ligand diversity and the averaging process in the occupancy calculation

These individual factors lead to a number of final conclusions. First, it appears to be unlikely for a single pseudoreceptor surface to fully and accurately replicate the individual protein binding pockets for a diverse ligand set. A low iso-level produces an iso-surface shell that contains the surfaces for all ligands, but can vary significantly from individual protein surfaces where there is diversity in a ligand set. Higher iso-level iso-surface shells more closely reproduce the surface that an "average" ligand would see, but lose the unique features of more diverse ligands. To address the drawbacks of the individual surfaces, it may be advantageous to use an ensemble of pseudoreceptor surfaces. RAPTOR implements a version of this with its dual-shell model[6]. In this model, two iso-surfaces are built using the most affine ligand as the basis for the inner shell and all ligands for the outer shell. We propose a similar solution, utilizing multiple shells of varied iso-levels: higher iso-level shells would represent the conserved portions of the ligand, and low iso-level shells would include the effects of ligand variation.

List of References

1. Tanrikulu Y, Schneider G (2008) Pseudoreceptor models in drug design: bridging ligand-and receptor-based virtual screening. Nat Rev Drug Discov 7:667–77.

2. Wilson G, Lill M (2011) Integrating structure-based and ligand-based approaches for computational drug design. Fut Med Chem 3:735–750.

3. Hahn M (1995) Receptor surface models. 1. Definition and construction. J Med Chem 38:2080–90

4. Hahn M, Rogers D (1995) Receptor surface models. 2. Application to quantitative structure-activity relationships studies. J Med Chem 38:2091–102.

5. Hahn M, Rogers D (1998) Receptor surface models. Perspect Drug Discov 12-14:117–133.

6. Lill MA, Vedani A, Dobler M (2004) Raptor: combining dual-shell representation, induced-fit simulation, and hydrophobicity scoring in receptor modeling: application toward the simulation of structurally diverse ligand sets. J. Med. Chem. 47:6174–6186.

7. Pei J, Zhou J, Xie G, et al. (2001) PARM: a practical utility for drug design. J Mol Graph Model 19:448–54, 472–3.

8. Pei J, Chen H, Liu Z, et al. (2005) Improving the quality of 3D-QSAR by using flexible-ligand receptor models. J Chem Info Model 45:1920–33.

9. Vedani A, Zbinden P (1998) Quasi-atomistic receptor modeling. A bridge between 3D QSAR and receptor fitting. Pharm Acta Helv 73:11–8.

10. Walters DE, Hinds RM (1994) Genetically Evolved Receptor Models. J Med Chem 37:2527–2536.

11. Chae CH, Yoo S-E, Shin W (2004) Novel receptor surface approach for 3D-QSAR: the weighted probe interaction energy method. J Chem Info Comp Sci 44:1774–87.

12. Berman H, Westbrook J (2000) The Protein Data Bank. Nucleic Acids Res 28:235–42.

13. Wilson GL, Lill MA (2012) Towards a realistic representation in surface-based pseudoreceptor modelling : a PDB-wide analysis of binding pockets. Mol Info 31:259-271.

14. Lorensen W, Cline H (1987) Marching cubes: A high resolution 3D surface construction algorithm. ACM Siggraph Comp Graph 21:163–169.

CHAPTER 4. INTEGRATED STRUCURE AND LIGAND-BASED METHOD FOR THE PREDICTION OF SITES OF METABOLISM OF CYTOCHROME P450 ISOZYMES.

***Note: This chapter was performed in collaboration with Dr. Laura Kingsley and Morgan Essex. Dr. Kingsley and Ms. Essex were responsible for the method development of the MD simulations, ensemble selection and generation, and docking, and performed these studies on CYP2C9. Gregory Wilson was responsible for the QSAR development and studies on CYP2C9. He also performed all studies for the remaining CYP isozymes and is responsible for webserver development. Portions of this chapter previously published in:

Kingsley, LJ.; Wilson, GL.; Essex, ME.; Lill, MA. Combining Structure- and Ligand-Based Approaches to Improve Site of Metabolism Prediction in CYP2C9 Substrates. *Pharm. Res.*, **2015**, *32*, 986-1001.

4.1 Introduction

As discussed in Chapter 1, our group has previously implemented a method that combined ensemble docking and NAT reactivity scores in an integrated structure and ligand-based tool for the prediction of CYP SoMs. The success of this approach was in part attributed to the inclusion of critical binding site conformations during docking via the use of a protein ensemble which led to a ~10% improvement in identifying reactive ligand poses as compared to docking to the crystal structure alone[1].

While the inclusion of protein flexibility using an ensemble of protein structures improved the generation of docking poses that were consistent with the experimentally known SoM, the number and diversity of false poses also increased. This increase in binding poses presents a significant challenge for the scoring functions used in docking and was thought to be the primary cause of the reduced prediction accuracy of docking observed in the top-1, top-2 and top-3 positions[1]. The poor docking performance in the ensemble is likely one of the key reasons that the improvement in SoM prediction accuracy in the ensemble was only modest compared to using only the crystal structure[1]. Based on our previous findings, we have developed a revised methodology to better incorporate protein flexibility and to better rank predicted poses in CYP2C9. The two main methodological improvements compared to our previous approach are a pre-filtering process to reduce the size of the protein ensemble used in docking and the implementation of pseudoreceptor modeling to accurately rank the binding poses relevant for SoM prediction. Compared to the existing methods cited above, our approach differs in method by which the data from docking and SMARTCyp[2] are combined, namely a modified pseudoreceptor scheme. To the best of our knowledge, this is the first attempt to directly incorporate SMARTCyp reactivity data into a pseudoreceptor model that is based on structural protein and ligand data to predict SoMs in CYP2C9.

A schematic of the revised procedure is shown in Figure 4.1. As with our previous model, both structure- and ligand-based principles were used in the current scheme; SMARTCyp, a successor of the NAT model was used to predict reactivity scores for each atom and ensemble docking was used to incorporate structural features of CYP2C9. We compared predictions in the crystal structure alone to predictions in a “pseudo-apo” ensemble which was selected based on a pre-filtering step used to isolate structures relevant for ligand binding. We found that incorporating “pseudo-apo” simulations increased the conformational space covered by the binding pocket allowing for successful docking of nearly all ligands. This was in stark comparison to the crystal structure alone, where less than 65% of the ligands tested could be successfully docked. In this study, as with our previous study, we noticed that the scoring function used in docking did not always accurately predict the correct binding poses. Due to the difficulty of accurately ranking poses using the docking scoring function, we introduced a pseudoreceptor model to differentiate between poses. Using the poses generated by docking and the reactivity scores generated by SMARTCyp, we produced a dataset suitable for pseudoreceptor modeling. A modified, in-house version of the RAPTOR[3] pseudoreceptor QSAR suite was used to develop a pseudo-receptor model to identify docking poses that correctly predict SoMs in our CYP2C9 ligand data set. With this approach we were able to significantly improve SoM prediction in the CYP2C9 ligand data set tested. Using a combination of docking to the pseudo-apo ensemble, SMARTCyp, and pseudoreceptor we were able to accurately predict the SoM in 96% of ligands within the top-2 positions.

Afterwards, we extended this method to eight additional CYP isozymes and obtained similar results. These isozymes are responsible for metabolism of significant percentages of all drugs: 1A2 (15%), 2A6 (3%), 2B6 (8%), 2C8 (8%), 2C19 (12%), 2-D6 (25%), 2E1 (4%), and 3A4 (50%), along with 2C9 at 20%[4].

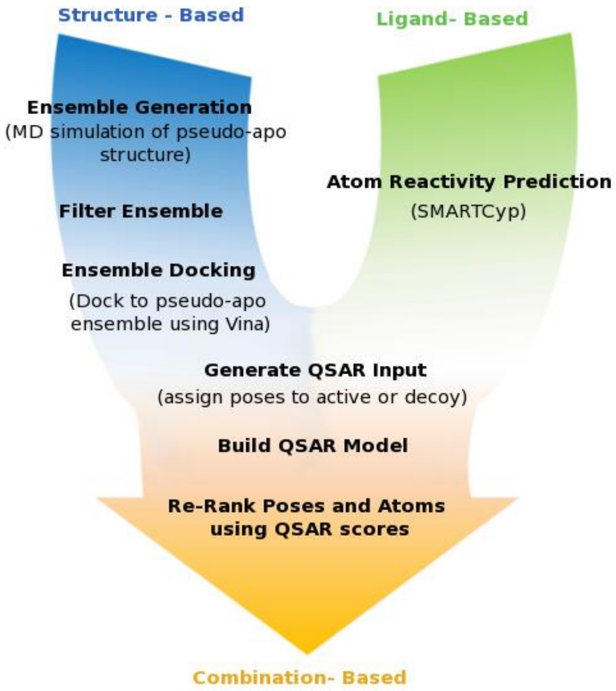

Structure - Based
Ligand- Based
Ensemble Generation
(MD simulation of pseudo-apo structure)
Filter Ensemble
Ensemble Docking
(Dock to pseudo-apo ensemble using Vina)
Atom Reactivity Prediction
(SMARTCyp)
Generate QSAR Input
(assign poses to active or decoy)
Build QSAR Model
Re-Rank Poses and Atoms using QSAR scores
Combination- Based

Figure 4.1: SoM prediction using a combination of structure- and ligand- based approaches. Using both atom reactivity data from SMARTCyp and structural data from docking, we generated a set of active (true SoM is within 4Å of the reactive oxygen) and decoy poses. A subset of these poses was used to train a pseudo-receptor QSAR model which was then used to evaluate all docking poses of all ligands.

4.2 Methods

4.2.1 CYP2C9 Ligand Library Preparation

A set of 73 structurally diverse CYP2C9 substrates with known SoMs were used for this study (Table A12). The compounds were based on those used by Danielson et. al(20), which were taken from the literature[5] and the University of Washington Metabolism and Transporter Drug Interaction database© (www.druginteractioninfo.org). All possible stereoisomers (in case that stereochemistry was not defined for the structure) and relevant protonation states were considered as unique chemical structures resulting in 139 total ligand structures. Ligands were built using Maestro and minimized using MacroModel as previously described[1].

4.2.2 SMARTCyp

SMARTCyp[2] is a reactivity model that predicts the reactivity at C, S, N, and P positions in a given ligand based on a series of over 40 rules derived from quantum calculations. SMARTCyp 2.4.2 was used to predict likely SoMs based on reactivity energies and atom accessibility in each of the 139 total ligand structures. The atoms of each ligand were then ranked according to the predicted abstraction energy, also referred to as the SMARTCyp score. In cases where one or more ligand variants existed, for instance two possible protonation states of the same ligand, the best (lowest) overall score was selected for each atom. The percentage of correctly predicted SoMs in the top-1, top-2, and top-3 positions were calculated using the experimentally known SoMs. In cases where a given substrate had more than one known SoM, only the highest predicted SoM was considered. This criterion was also used for all subsequently described methods.

4.2.3. Static Crystal Docking

The 1R9O crystal structure of CYP2C9 was used for the static docking studies. The co-crystalized ligands, flurbiprofen and glycerol, along with all crystal waters were removed. The crystallized heme (deoxygenated) was replaced by an oxygenated heme. Protonation and tautomer states of histidine and rotamer states of asparagine, glutamine and histidine were assigned using Reduce[6]. The ligand library was prepared for docking using the PyMol plugin developed by Danielson et.al[7].

4.2.3.1 Autodock Vina

Ligands were docked using AutoDock Vina (Vina). The docking volume was defined using our PyMol plugin. The selected docking cuboid was roughly 20Å on each side and included the active face of the heme and surrounding residues that could be relevant for binding. Default values were used for all docking parameters in Vina. For each unique ligand, 10 total docking poses were generated and 5kcal/mol was chosen as the maximum energy difference allowed between the best and any other reported docking pose.

4.2.3.2. Ranking

Docking success was evaluated based on the distance between the known SoM and the reactive oxygen of the heme moiety. Because docking to multiple similar protein structures can result in redundancy of several ligand poses, the poses were clustered using k-medoid clustering. K was iteratively adjusted such that the maximum RMSD between any two poses and the cluster center was less than 1.0Å. The pose with the best (lowest) docking score from each cluster was selected as the representative member for that cluster.

Next, docking poses of all protonation states and stereoisomers of a given ligand were pooled resulting in a single set of poses for each ligand containing all protonation and tautomeric states of the ligand. The combined poses were then ranked according to the docking score. If two poses had the same docking score, both would share the same rank, but the rank immediately following would reflect the inclusion of multiple poses. For instance assuming two poses had the same score and were ranked first, the next compound would be ranked in the third position to account for the two ligands that had been previously ranked higher.

A distance of 4.0Å or less between any heavy atom and the reactive oxygen was considered to be potentially reactive. Poses that did not have a heavy atom within 4.0Å of the reactive oxygen were omitted from the ranking scheme.

Next, each atom was assigned the best Vina docking score attained by any pose wherein the atom was within 4.0Å of the oxygen on the heme. The atoms were then ranked according to the assigned score and the percentage of accurately predicted SoMs that occurred in the top-1, top-2 and top-3 positions was calculated. In addition to determining the accuracy of SoM predictions in the top-3 positions, the overall docking success was determined for docking to the crystal structure and the ensemble. The overall docking success is defined as the percentage of ligands that could be successfully docked regardless of ranking. In other words, the overall docking success is a measure of how well the pose prediction portion of the docking algorithm performed exclusive of the scoring function.

4.2.4 Ensemble Generation

4.2.4.1 MD Simulations

An MD simulation of the pseudo-apo structure of CYP2C9 was used in the ensemble generation process. To generate the initial protein structure the ligand was removed from the CYP2C9 crystal structure, 1R9O.

The MD simulation was performed using Gromacs-4.5.5[8,9] and the Amber03 force field. The input structure was prepared using Reduce[6] to identify the proper rotamer, tautomer, and protonation states of histidine, and the proper rotamer states of asparagine and glutamine. The heme parameters were extracted from the literature[10]. We opted to use an oxygenated heme model because the oxygen may be critical for the docking of certain ligands. Gromacs was used to solvate the system in an octagonal water box of SPC216 waters and 6 chlorine ions were added to neutralize the system. The box size was selected to guarantee a minimum distance of 15Å between solute and box edge.

The steepest descent method and particle mesh Ewald (PME) summation with a grid size of 0.12nm was used to carry out 1000 steps of energy minimization. To compute van der Waals interactions a switching function was applied between 1.0nm and the cut-off of 1.4nm. The LINCS algorithm[11] was used to constrain bonds containing hydrogen atoms. Next the hydrogen bond network of the surrounding waters was established using a 200ps MD simulation in which all but the waters were restrained. Simulations were performed at 300K using PME, Berendsen thermostat, and Parrinello-Rahman pressure coupling. The integration time step was 2fs. Finally a 400ps equilibration run was performed to equilibrate the system prior to the 10ns production run.

4.2.4.2. Ensemble Generation and Refinement

The initial ensemble was generated by extracting frames every 100ps from the pseudo-apo production run. The initial ensemble was then refined using a docking-based filtering process resulting in a 6 member ensemble. From the 73 ligands used in this study, 14 structurally diverse ligands (denoted in Table A12) were manually selected for ensemble refinement. These 14 ligands were docked to all members of the 100-member ensemble using AutoDock Vina, as described above. Ligand variants were combined to give a single set of poses for each ligand as previously described.

To determine which protein structures were able to best dock the 14-ligand training set, a fitness score was calculated for each protein structure as follows:

$$Fitness = \frac{\sum_{i=1}^{14-ligand\ training\ set} w_i f_i}{\sum_{i=1}^{14-ligand\ training\ set} w_i} \qquad f_i = \left\{ \begin{matrix} 5; & rank = 1 \\ 4; & 2 < rank < 4 \\ 3; & 4 < rank < 6 \\ 2; & 6 < rank < 10 \\ 1; & rank > 10 \end{matrix} \right\} \qquad (4.1)$$

Where w_i is one over the number of protein structures to which the ligand i was successfully docked and f_i represents an assigned value based on the docking rank. Poses that were not successfully docked (e.g. did not have the known SoM within 4Å of the reactive oxygen) were given a score of 0, while those that were successfully docked were given a score between 1 and 5, based on the rank of the pose. The factor w_i guarantees that protein structures are more likely selected for the refined ensemble that allow the successful docking of ligands that are difficult to dock. For example, assume that two ligands A and B dock successfully to protein structure S. Assume ligand B is successfully docked to 49 other protein structures (out of 100 structures in the initial ensemble) and ligand A is only docked successfully to S. As protein structure S seems to be unique and relevant for binding ligand A and structurally similar ligands, it should gain a high fitness value and be more likely to be included in the refined ensemble. This is achieved by the introduction of the weight (w_i) which will be 1.0 (1/1) for ligand A but only 0.02 (1/50) for ligand B.

The protein structures from the ensemble were then ranked by fitness. We found that 13 out of 14 ligands could be successfully docked to at least one of the top-5 ranked protein structures. The remaining ligand, 2-oxoquazepam, was not successfully docked until the 34th ranked protein structure. Based on our previous findings that the inclusion of multiple protein conformations can be problematic for the docking scoring function, we felt that including 34 structures would be detrimental to the model. We tested the top-4, top-6, top-8, and top-10 protein structures on the entire ligand data set (data not shown) and found that selecting the top-6 structures achieved optimal template diversity.

4.2.5. Ensemble Docking

All 139 ligand structures were docked to the 6-member protein ensemble. Ensemble docking was performed in a similar fashion as to the static crystal docking described above. Again, all ligand variants from all ensemble members were pooled to produce a single set of poses for each ligand, the resultant poses were clustered and the cluster member with the highest docking score was selected.

4.2.6. Ranking

As with the static crystal docking, the atoms of each ligand were ranked according to the best docking score in which that atom was within 4.0Å of the reactive oxygen. The percentage of successfully predicted SoMs in the top-1, top-2 and top-3 positions were calculated as well as the overall docking success, as described above.

4.2.7. SMARTCyp + Docking

In an attempt to improve SoM ranking in the top-1, top-2, and top-3 positions, we combined SMARTCyp reactivity predictions with the docking results. A single combined score (*CS*) was calculated for each atom of a given ligand using the following function:

$$CS = R_i + \gamma S_i \quad (4.2)$$

where R_i is the atom's SMARTCyp reactivity score (usually ranging from about 50 (best) to 100(worst)) and S_i is the docking score from the highest ranked pose where the atom *i* was within the 4.0Å cutoff from the oxygen of the heme (usually ranging from about -12 (best) to -6 (worst)). Gamma (γ) is a weighting factor between 0 and 10, and is used to adjust the contribution of the docking score (S_i) to the total combined score (*CS*). In order to be further considered in the *CS* scheme, an atom had to have both a docking score and a SMARTCyp score, otherwise the atom was omitted as a potential SoM.

Gamma was optimized using a subset of ligands (denoted in Table A12) and the following fitness function:

$$fitness = (\%top1) + 0.5(\%top2) + 0.25(\%top3) \quad (4.3)$$

Where *%top1*, *%top2*, and *%top3*, reflect the percentage of accurately predicted SoMs in the top-1, top-2 and top-3 positions, respectively. Gamma was initially set to 0 and was iteratively increased by 0.5 to a maximum of 100. The gamma value that maximized the fitness score for each data set (*i.e.* crystal or pseudo-apo ensemble) was selected.

For each ligand, atoms were ranked by *CS* value. As with the docking scores, atoms with equivalent *CS* values were ranked at the same position, but the next position reflected the inclusion of multiple atoms at the previous position. The percentage of correctly identified SoMs in the top-1, top-2 and top-3 ranked atoms was calculated for the x-ray crystal structure alone and the pseudo-apo ensemble.

4.2.8. SMARTCyp+ Docking+ QSAR

In an attempt to further improve SoM prediction results we implemented a modified QSAR scheme to evaluate and re-rank docking poses. The SMARTCyp score and free energy of binding were combined into the fitness functions used for deriving the QSAR model.

4.2.8.1 Dataset Preparation and Selection

As described previously, SMARTCyp assigns reactivity scores to all ligand atoms, with the lowest score representing the predicted SoM. When combining the SMARTCyp reactivity approach with docking, the SoM predictions can be re-ranked by including only those atoms within a reactive distance of the oxygen atom of the heme. The main limitation of this approach is the accuracy of the docking scoring function. Often poses are found in which the true SoM is within the cutoff distance (active poses), but these poses may be amongst the worst ranked by the scoring function. This problem intensifies as more poses are introduced using ensemble docking. To overcome the limitations of docking scoring functions, we developed a modified version of the RAPTOR QSAR package to generate a statistical model to differentiate poses that are consistent with the experimentally known SoM from those which are not.

The clustered docking poses were used as input for the QSAR model (Figure 4.2a). The poses were first separated into active poses and decoy poses (Figure 4.2b**)**. An active pose was defined as a pose in which the known SoM was docked within 4Å of the oxygen on the heme and had the known SoM within the top-3 ranked SMARTCyp scores for those atoms within 4Å of the heme oxygen atom. The active poses were further classified by whether the known SoM had the first, second, or third best SMARTCyp score **(**Figure 4.2c**)**. A decoy pose was defined as any pose that was docked with at least one atom within 4Å of the heme oxygen atom but did not meet the criteria for an active pose.

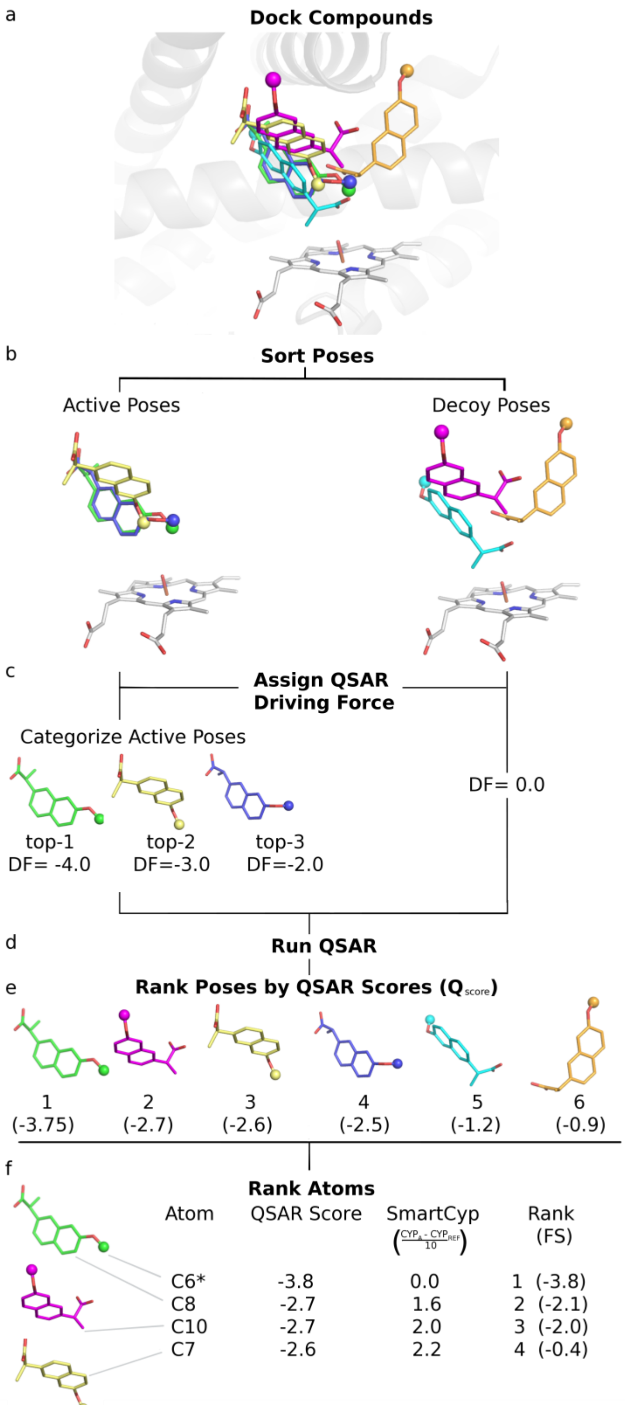


Figure 4.2: Scheme of QSAR modeling process. First the poses generated by docking (**a**) were separated into active and decoy poses and the actives subcategorized into Top 1, Top 2, and Top 3 actives(**b**). A driving force (DF) was then assigned to each pose (**c**) and the RAPTOR package was used to generate a QSAR model(**d**). After the QSAR training, all poses for a ligand are sorted by the QSAR score (**e**) and atom scores are assigned to the top three SMARTCyp atoms for each pose. Atoms are then ranked by the final score (FS) according to the QSAR model (**f**).

4.2.8.2. Test and Training Selection

A random set of nineteen ligands was selected as the initial test set for the QSAR simulations and the remaining ligands were assigned to the training set. The value of nineteen was chosen as this represented approximately one-quarter of the available ligands for QSAR modelling. The test set was then manually curated to ensure that it covered the chemical space of the training set. During this evaluation, four of the test ligands were moved to the training set, and an equal number of ligands were moved to the test set to retain the overall 3:1 training to test ratio. Two of the ligands that were moved into the training set had unique ring structures not found in any other ligand in the data set, a third ligand had a unique long carbon chain, and the final ligand was the smallest compound in the data set. These unique features cause the ligands to be unsuitable for the test set. This test set was then used for all remaining QSAR simulations. The final training and test sets are noted in the "Data Set" column of Table A12.

As discussed earlier, active poses are further classified based on the rank of the SoM using the SMARTCyp score. Thus, for many ligands there are binding poses in which the SoM is ranked as most reactive atom (i.e. other more reactive atoms are not within 4Å of the catalytic center) and other poses where the SoM is ranked lower (e.g. as top-2 or top-3) because in addition to the try SoM, other more reactive atoms also fall within 4Å of the reactive oxygen. In a strict sense, the later poses disagree with the experimental SoM data and would add noise to the QSAR training process. Thus, during QSAR model training only the active poses with the highest ranked SoM based on the SMARTCyp score were used as active poses. All other active poses, however, were moved into the final prediction set which contains all docked poses with any atom within 4Å of the catalytic oxygen. This prediction set was used for final evaluation of SoM prediction quality using our optimized QSAR model (Figure 4.1, last step).

4.2.8.3 Inclusion of SMARTCyp reaction scores

To directly incorporate the SMARTCyp scores into the QSAR model, the RAPTOR package was modified. The original version of RAPTOR uses hydrogen-bond interactions and hydrophobic contacts between the ligands and the pseudo-receptor generated by RAPTOR to predict binding affinities. In the modified version of RAPTOR, the SMARTCyp score was provided as an additional contribution to the overall predicted score. Thus, the QSAR score Q_{score} was computed by the sum of hydrogen-bond interactions ΔG_{HBond}, hydrophobic contacts ΔG_{HPhob} and SMARTCyp score $S_{SMARTCyp}$:

$$Q_{score} = \Delta G_{HBond} + \Delta G_{HPhob} + 0.1 \cdot S_{SMARTCyp} \qquad (4.4)$$

SMARTCyp scores were assigned to every pose as 1/10th of the original value to scale the reactivity scores to the same order of magnitude as the other two contributions to the Q_{score} within the RAPTOR models. For active poses, the SMARTCyp score of the known SoM was used. For decoys, the lowest SMARTCyp score of any atom within 4.0Å of the oxygen atom of the heme was used.

Also, the input to the QSAR method was adjusted **(**Figure 4.2b**)**. Typically, all poses for a given ligand are treated as alternative conformations of the same ligand and the experimental affinity value is used during the QSAR modeling process for every conformation. For our method, we grouped active and decoy poses separately. In addition, instead of binding affinities, the active poses were assigned a negative score, while the decoys were assigned a score of 0 or a positive value. We will refer to the difference between these scores as a “driving force.” The goal of this driving force is to identify the physicochemical features in the QSAR model that allows discrimination between active and decoy poses due to differences in protein-ligand interactions.

In order to determine the optimal driving force, we ran multiple RAPTOR simulations with different driving forces. We ran simulations with both a fixed driving force for all active poses, and simulations with a variable driving force for the actives. For the variable driving force simulations, the top-1 actives poses are assigned a value of X-Y, top-2 poses are scored as X, and top-3 poses are scored as X+Y where X and Y are real floating point values ranging from -5 to -2 and Y ranging from -1.5 to -0.5. Using variable weights for top-1, top-2 and top-3 poses improved the performance of the QSAR model compared to assigning identical weights to all actives. Many of the driving force weights generated QSAR models with similar quality. Therefore, we chose a set of weights in the middle of our testing range, i.e. an X value of -3 and a Y value of 1, with the decoy set being assigned a value of zero. This setting had the best performance by a slight margin.

4.2.8.4 QSAR Modeling

The modified RAPTOR program was used to generate a pseudo-receptor QSAR model for CYP2C9 with all remaining parameters set to their default values. Five individual models, run with the fast search mode, a coupling factor of 0.5 and sharpness penalty of 1 were constructed for each modeling run.

4.2.8.5 Analysis of QSAR Results

Typically, pseudo-receptor models are used to predict the binding affinity of a ligand. RAPTOR, in addition to providing an overall prediction of the affinity of the ligand, predicts the binding energy for each conformation of a ligand. In this study, those conformations are the individual docking poses for a ligand. However, here the predicted score does not provide an estimate of the binding affinity but yields a likelihood score for each conformation to be the pose predicted to have the known SoM within 4 Å of the catalytic center. To evaluate the success of our model, all binding poses of training and test set were combined with the predictive set of actives excluded from the modeling process. The trained QSAR model was used to assign QSAR scores to all poses which were then ranked by this score (Figure 4.2e). The atoms within 4Å of the catalytic center with the top-3 SMARTCyp scores for each pose were assigned modified QSAR scores using the following formula:

$$FS_A = Q_{score} + \frac{CYP_A - CYP_{REF}}{10} \qquad (4.5)$$

where FS_A is the final score for atom A, Q_{score} is the QSAR score for the pose in which atom A is found, CYP_A is the SMARTCyp score for atom A, and CYP_{REF} is the SMARTCyp score for the atom used in the QSAR model building process . This formula adjusts the QSAR score for the difference in SMARTCyp scores between the top three SMARTCyp atoms **(**Figure 4.2f). The lowest score for any given atom among all poses was identified, and then the atoms themselves were sorted by score. The highest ranked known SoM was identified and the percentage of correctly predicted SoMs in the top-1, top-2, and top-3 positions were reported.

4.3 CYP2C9 Results and Discussion

4.3.1 SMARTCyp Prediction

Several reactivity schemes have been developed to predict SoMs in CYP substrates based on the physicochemical properties of the ligand alone[2,12]. Such ligand-based methods are advantageous because they do not require protein structural information and are computationally efficient. SMARTCyp is one example of a widely used reactivity based method. Potential SoMs are evaluated based on a combination the accessibility of the atom within the structure and the estimated energy required to abstract a hydrogen from carbon atoms or for an oxygen attack in the case of nitrogen, phosphorus, and sulfur atoms. The resultant score is referred to as a SMARTCyp score and is used to rank potential SoMs. Recently, a new version of the SMARTCyp program, version 2.4.2, was released with parameters specific to CYP2C9 ligands[13].

We generated a 139-ligand data set comprised of all possible rotameric and protonation states of 73 unique ligands and evaluated each using SMARTCyp version 2.4.2 (referred to as SMARTCyp). The atoms of each ligand were ranked according to the assigned SMARTCyp reactivity score and the number of correctly predicted SoMs in the top-1, top-2, and top-3 positions was calculated (Table 4.2 - SMARTCyp Alone column). SMARTCyp correctly predicted the known SoM at the top-1 position in 42% of the ligands tested. In the top-2 and top-3 positions, the prediction percentages increase to 58% and 67% respectively.

4.3.2. Static Docking

Docking is another approach used to predict potential SoMs in CYP ligands. Docking is one of the most widely used techniques in structure-based drug design and provides information about potential ligand binding modes. In the biologically active conformation within the CYP binding site, the ligand should be positioned in such a way that the SoM is in close proximity to the reactive oxygen atom of the heme moiety. In theory, if the docking pose is correctly predicted, atoms which are positioned near the oxygen atom of the heme are the most likely SoM candidates.

As a comparison to our new approach, we docked our ligand library into the crystal structure of CYP2C9 (PDB ID: 1R9O) using Autodock Vina (Vina). A docking pose was considered to be an accurate SoM prediction if the distance between the known SoM and the reactive oxygen of the heme moiety was 4Å or less. Docking poses were ranked by the internal Vina scoring function and the percentage of correctly predicted SoMs in the top-1, top-2, and top-3 ranked poses were calculated (Table 4.2 - Vina Alone column). In addition to assessing predictions in the top-3 ranked poses, we calculated the overall docking success by determining the percentage of ligands that achieved an active pose regardless of rank (Table 4.2- Vina Alone column).

SMARTCyp outperformed docking in identifying the known SoM within the top-3 positions. However, the overall docking success was approximately equal to the prediction success of SMARTCyp in the top-3 positions (64% and 67% respectively). This highlights two possible shortcomings in the standard Vina docking approach. First, despite 67% overall accuracy in docking, less than half of these poses were ranked in the top-3 positions, suggesting that the Vina scoring function does not always rank potentially biologically active conformations in the top positions. Second, the failure to achieve 100% docking success suggests that the binding pocket of the crystal structure alone may not be able to accommodate the structural diversity of the ligands in the data set.

It is well known that CYP enzymes are highly flexible and that the binding sites of these enzymes often have to adapt to accommodate structurally different ligands[14]. The plasticity of the CYP binding sites can make docking to these enzymes challenging and often ensemble approaches are employed to improve docking results[15,16].

4.3.3 Ensemble Generation and Selection

4.3.3.1 Ensemble Diversity

A pseudo-apo ensemble was generated by extracting 100 snapshots from a 10ns trajectory of CYP2C9 with the crystal ligand removed. A principle component analysis (PCA) suggests that through the duration of the simulation, both the overall protein structure and the binding site residues adopted several alternative conformations (Figures 4.3a and 4.3b). Ultimately, the increased diversity in the pseudo-apo ensemble allowed for improved docking of several ligands in comparison to the crystal structure.

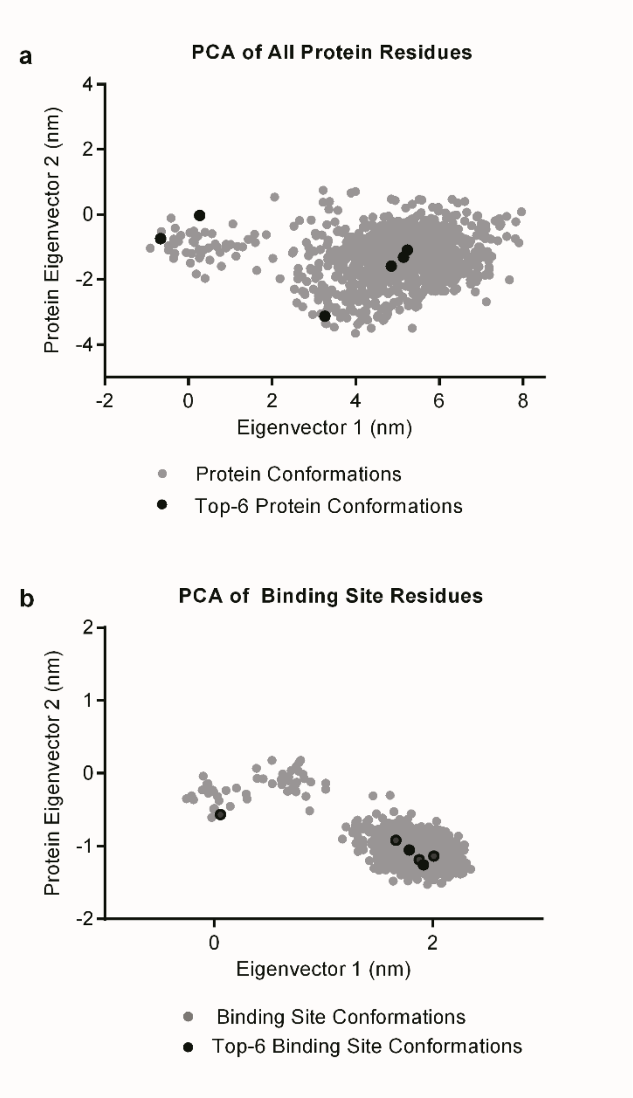


Figure 4.3: Principal component analysis of CYP2C9 pseudo-apo MD trajectories. The PCA using all protein residues **(a)** and only the binding site residues (**b**). The binding site residues were manually defined based on the defined binding site box from the docking simulations. The black circles represent the top-6 structures selected for the final ensemble.

For instance, no active docking pose was found for 9-cis-retinoic acid in the crystal structure, however side-chain rotations that occurred during the pseudo-apo simulation allowed for successful docking of this ligand (Figure 4.4a). The orientation of LEU 208, PHE 476, and PHE 100 (PHE 100 was omitted from figure for clarity) are crucial to achieve a bioactive conformation of this ligand. In the crystal structure, the top-ranked bioactive pose of the ligand directly overlaps with LEU 208. Furthermore, the rotation of ASP 293 in the pseudo-apo simulation provides a potential hydrogen bonding site for the ligand.

Additionally, the binding of torsemide required a significant rearrangement of residues in the active site (Figure 4.4b). A ~3Å shift in the C-terminal loop is accompanied by the ~180 degree rotation of PHE476 in the pseudo-apo simulation which allows for this ligand to be successfully docked. In the closest-to-active pose in the crystal structure docking the ligand is found to occupy a pocket created between the C-terminal loop and the G helix, resulting in a conformation where the SoM is 4.2Å from the reactive oxygen. In the pseudo-apo simulation, shifting of the C-terminal loop causes a closure of this pocket and causes the ligand to bind on the opposite side of the C-terminal loop where the SoM is within 3.7A of the reactive oxygen and at a more favorable angle to the reactive oxygen.

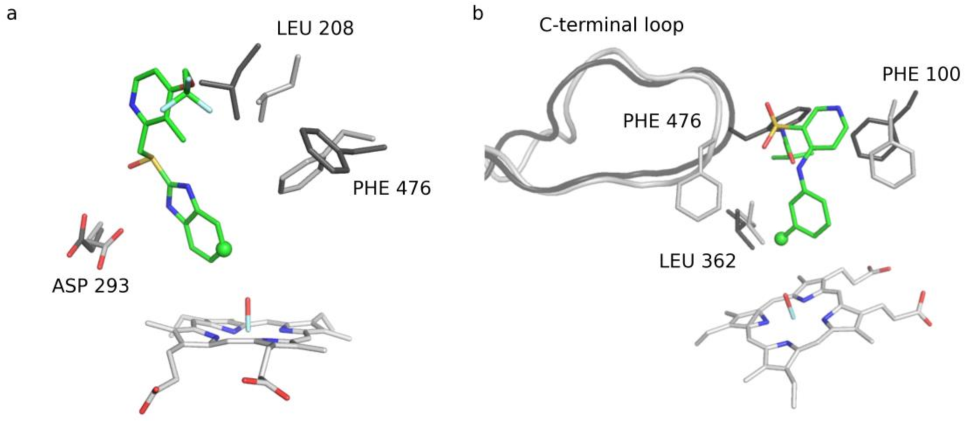


Figure 4.4: Conformational adaptation in the pseudo-apo simulation that allows for successful docking of 9-cis-retinoic acid and torsemide. The true SoM for each ligand has been denoted with a sphere. Compared to the flurbiprofen-bound crystal structure (shown in dark grey sticks and cartoon), several residues and the C-terminal loop adapt to allow for ligand binding. In the case of 9-cis-retanoic acid **(a)**, LEU208 and PHE 476 rotate to allow for the ligand to fit into a bioactive conformation. Furthermore ASP293 rotates into a position to allow a potential hydrogen bond to the imidazole ring. In the case of torsemide **(b)**, a ~180º rotation of the side chain and a >3Å shift in the C-alpha position of PHE 476 was observed, allowing for a bioactive conformation of torsemide that was not observed in the crystal structure. This shift closes a pocket between the C-terminus and the G helix (not shown). The closest-to-active pose found in the crystal structure docking was found to occupy this pocket. Closure of this pocket allows for an alternative ligand conformation to be found in the pseudo-apo ensemble that is consistent with the known SoM of torsemide.

4.3.3.2 Final Ensemble Selection

While, the inclusion of a variety of binding site conformations may be essential for docking of large and diverse ligand libraries such as the one tested here, an ensemble of several hundred members is both cumbersome and redundant; therefore the pseudo-apo ensemble was further refined.

A docking filter was used to select the most relevant conformations from the initial ensembles. Using a subset of 14 ligands and the fitness function described in the Methods section, the top-6 structures from the pseudo-apo ensemble were selected as the final ensemble members. The fitness scores, RMSD to the crystal structure, as well as the binding site volume are shown for each member of the ensemble in Table 4.1.

Table 4.1: Calculated fitness score, overall RMSD to 1R9O crystal structure and binding site volume of selected ensemble members. The volume of the binding site over the course of the trajectory calculated using POVME

Structure[a]	**Fitness Score**	**RMSD to Crystal**	**Binding Site Volume**
PA 97	2.66	1.18	361
PA 19	2.58	1.19	393
PA 66	2.58	1.02	422
PA 1	2.55	1.12	836
PA 25	2.53	1.18	363
PA 91	2.51	1.08	411

The 14-ligand training set was initially docked to all protein structures (Figure A1). Although some individual members of the ensemble perform worse than the crystal structure alone, taken together, ensemble docking shows significant improvement over docking to the crystal structure alone. The crystal structure successfully docked only half of the 14-ligand test set whereas the pseudo-apo ensemble docked successfully 13 out of 14 compounds into the top-5 protein structures alone (Figure 4.5). The remaining ligand, 2-oxoquazepam was successfully docked to the 34th ranked structure.

The selected ensemble members were found to be structurally diverse and to cover a significant portion of the conformational space sampled by the MD simulation according to the PCA (Figure 4.3). Compared to randomly selected ensembles of the same size, the filtered ensemble provides considerable improvement in the docking results in the top-3 positions and slight improvement in the overall prediction success (Table 4.2). The improvement over random selection indicates that the pre-filtering procedure aids in the isolation of protein conformations relevant for docking.

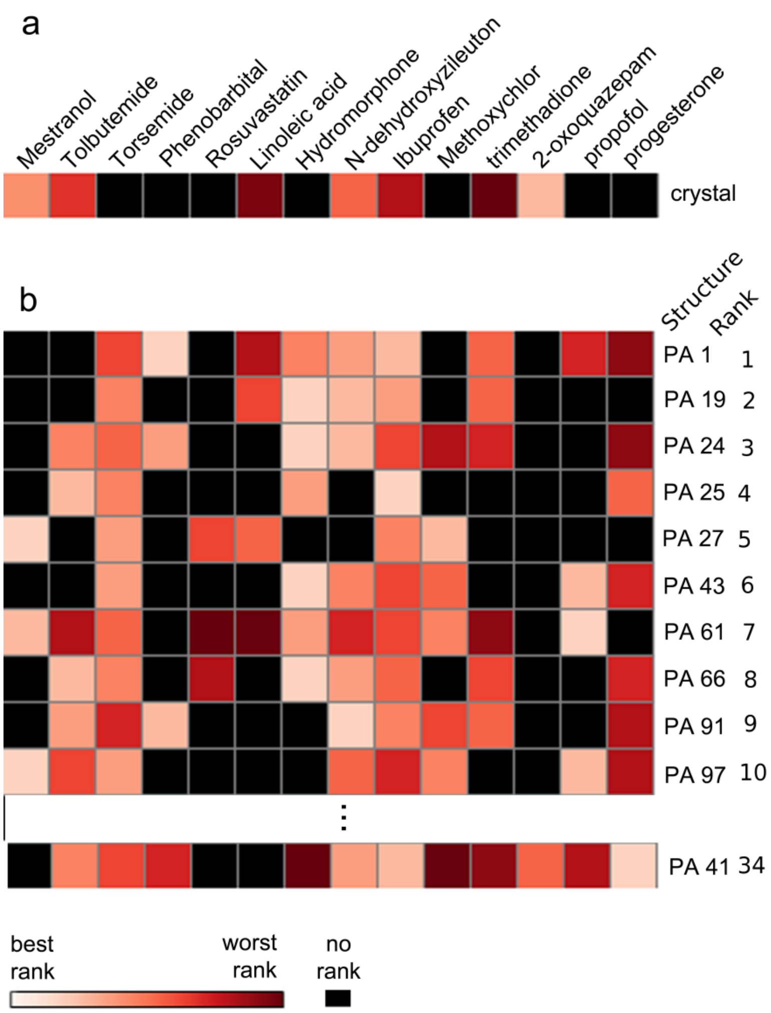


Figure 4.5: A visual representation of the docking performance in a) the crystal structure, **b)** the pseudo- ensemble using the 14-ligand subset. Ligand ranking is indicated by the shade of red, lighter regions indicate highly ranked poses, while black indicates that no pose was found in which the true SoM was within a reactive distance to the oxygen on the heme. The rank of the protein structure according to the fitness function is shown on the far right hand side for the pseudo-apo ensemble.

4.3.4 Ensemble Docking

Following the selection of the 6-member pseudo-apo ensemble, we used Vina to dock all 73 compounds in the data set to each member of the ensemble. Compared to the crystal structure alone, the ensemble offered significant improvement in the top-1, top-2 and top-3 positions and in the overall docking success (Table 4.2).

The more diverse binding pockets of the mixed ensemble are likely responsible for the significant improvement in the overall docking success. The increase in accurate predictions in the top-1, top-2 and top-3 positions, while significant, does not match the improvement in overall docking success. The increased binding pocket diversity in the ensemble is likely the reason that more compounds can be successfully docked, however, this diversity can also result in a higher number of alternative ligand poses, making the identification of true positive poses more challenging for the docking scoring function. This is one possible reason that the individual increase in the top-1, top-2, and top-3 positions is not as drastic as in the overall docking performance.

Table 4.2: Comparison of various methods for predicting SoMs in the top-1, top-2, and top-3 positions.

			X-ray structure alone[h]		
	Random[a]	SMARTCyp Alone[b]	Vina Alone[c]	Vina + SMARTCyp[d]	Vina+ SMARTCyp + QSAR[e]
Top-1	12%	44%	21%	38%	49%
Top-2	24%	59%	27%	53%	56%
Top-3	38%	68%	37 %	60%	63%
% docked			64%	Gamma 0.0[g]	
			Pseudo-apo Ensemble[i]		
			Vina Alone[f]	Vina + SMARTCyp	Vina+ SMARTCyp+ QSAR
Top-1			44% [28 ±6.5%]	55%	88%
Top-2			58% [37 ±5.7%]	77%	96%
Top-3			67% [48± 7.0%]	88%	96%
% docked			96% [92±1.7%]	Gamma 23.5	

[a] Percentage of correctly predicted SoMs if a heavy atom was chosen at random for each ligand.

[b] Percentage of correctly predicted SoMs using SMARTCyp only.

[c] Percentage of correctly predicted SoMs using Autodock Vina alone. A prediction was considered "correct" if the true SoM was within 4.0Å in the top-1, top-2 or top-3 ranked docking poses, respectively.

[d] Percentage of correctly predicted SoMs using a combination score comprised of the Vina score and the SmartCyp score, see Methods section for full details.

[e] Percentage of correctly predicted SoMs using the modified QSAR model that includes the poses provided by Vina docking and the reactivity scores from SMARTCyp.

[f] Bracketed values represent the percentage of successfully docked compounds when the protein structures that comprised the ensembles were chosen at random. These values represent the average and standard deviation over three randomly selected protein sets.

[g] Although a gamma (γ) of 0 is selected, the omission of some atoms due to failure to find both a successful docking pose and SMARTCyp score can result in slightly different rankings using the *CS* versus SMARTCyp. These differences were caused by the inability to find a successful docking pose, therefore an atom may be ranked in SMARTCyp but not the combination

approach, which can result in slight changes in the overall rankings as observed in the crystal structure.

[h] Binding poses were identified using docking with AutoDock Vina to the x-ray structure of CYP2C9 only.

[i] Binding poses were identified using docking with AutoDock Vina to an ensemble of proteins structures generated by an MD simulations based on the pseudo-apo form of CYP2C9 (holo x-ray structure with co-crystallized ligand removed).

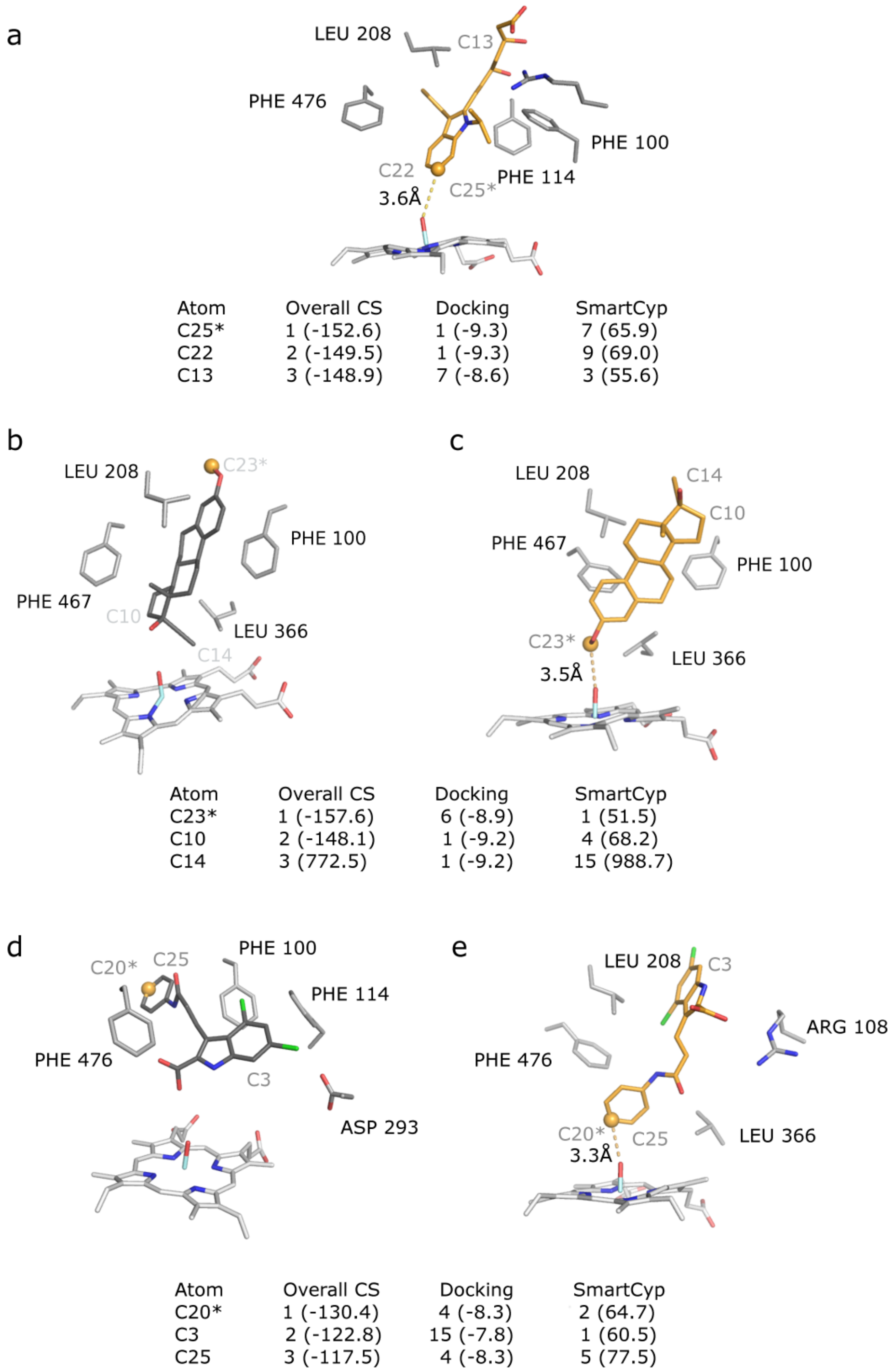
4.3.5 Combining Docking and SMARTCyp

We hypothesized that SoM predictions could be further improved by combining the structural data from docking and the ligand-based reactivity predictions from SMARTCyp. For instance, let us assume that SMARTCyp incorrectly predicts a given atom as the true SoM; although the incorrectly predicted atom may be a highly reactive, it may not be a structurally feasible SoM based on its binding conformation. For example, it may be part of a bulky group that cannot easily fit close to the reactive oxygen of the heme. By including contributions from both docking and SMARTCyp, such atoms could be re-ranked or even eliminated as possible SoMs, resulting in improved predictions.

The optimized gamma value can offer insight about the individual contributions of docking scoring and SMARTCyp to the overall ranking of the SoM; a low gamma suggests that SMARTCyp dominates the calculated *CS* and docking only provides a minor contribution, a gamma of around 10 would suggest approximately equal contributions of both docking and SMARTCyp, and a large gamma would suggest that docking dominates the *CS* function.

In the crystal structure, the optimized gamma value of 0.0 suggests that the results are entirely dominated by the SMARTCyp rankings of the compounds. On the other hand, the gamma score for the pseudo-apo ensemble is 23.5, suggesting that docking scores are a major contribution to the overall *CS* ranking. There are several reasons for this discrepancy in gamma scores. Most notably, in the pseudo-ensemble the rankings of SMARTCyp and docking are approximately equal in the top-3 positions (~65%). This suggests that both docking and SMARTCyp have approximately equal ability to contribute to the final ranking. However, almost all compounds can be successfully docked to the ensemble, indicating that docking has the potential to further improve SoM prediction above the ~65% observed with either approach individually. This is in contrast to the crystal structure where docking has a significantly lower percentage of compounds ranked in the top-3 (~45%) and also a lower overall docking success (~65%), thus a less significant potential to contribute to the overall *CS* ranking.

In Figure 4.6, we provide some specific examples of how *CS* ranking in the pseudo-apo ensemble improved SoM prediction in various compounds. In some compounds, such as galangin, the contribution of the docking score was essential for the top-1 *CS* ranking of the compounds (Figure 4.6a), whereas in others, like terbinafine, it was the SMARTCyp (Figure 4.6b) score that was the determining factor. SMARTCyp and docking did not rank the same ligands in the top-3 positions as was seen with galangin and terbinafine. These differences allowed for approximately 10% improvement in the *CS* ranking. However, the most intriguing cases were those in which different rankings of individual atoms by SMARTCyp and docking led to a synergistic ranking effect. In these cases, the *CS* ranking of the true SoM was higher than in either approach individually. Suprofen, for example shows this trend (Figure 4.6c). Suprofen and other ligands where there was a synergistic effect accounted for the remaining ~10% improvement in the *CS* ranking is as compared to either method alone.



Atom	Overall CS	Docking	SmartCyp
C25*	1 (-152.6)	1 (-9.3)	7 (65.9)
C22	2 (-149.5)	1 (-9.3)	9 (69.0)
C13	3 (-148.9)	7 (-8.6)	3 (55.6)

Atom	Overall CS	Docking	SmartCyp
C23*	1 (-157.6)	6 (-8.9)	1 (51.5)
C10	2 (-148.1)	1 (-9.2)	4 (68.2)
C14	3 (772.5)	1 (-9.2)	15 (988.7)

Atom	Overall CS	Docking	SmartCyp
C20*	1 (-130.4)	4 (-8.3)	2 (64.7)
C3	2 (-122.8)	15 (-7.8)	1 (60.5)
C25	3 (-117.5)	4 (-8.3)	5 (77.5)

Figure 4.6: The combined score (*CS)* versus docking and SMARTCyp scores individually of the top-3 atoms as ranked by the *CS*. The top-ranked docking pose is shown on the left of each panel and the bioactive pose is shown in orange on the right. In panel **a**, the top-

ranked pose is the bioactive pose, thus only a single pose is shown. True SoMs have been denoted in the text with a star and in the structures as an orange sphere. **a)** In some cases the docking score is the determining factor for the overall CS ranking of the true SoM. In fluvistatin, for instance, in the top ranked docking pose the true SoM, C25, was the atom nearest to the reactive oxygen of the heme. Even when combined with a poor SMARTCyp score, the favorable docking score of this pose allowed this atom to be ranked 1st overall. In other cases, SMARTCyp is crucial for the ranking using *CS*. In the case of mestranol, the top-ranked docking pose places C10 and C14 nearest to the reactive oxygen (**b**). However, these atoms were ranked poorly by SMARTCyp (4th and 5th, respectively). The top-ranked bioactive pose (**c**) had a docking score that was only slightly less favorable than the top ranked pose, and thus when combined with the SMARTCyp scores, wherein the true SoM was ranked first, the overall *CS* ranking placed the true SoM in the top-1 position. In some cases there was a synergistic outcome using the *CS*. In GV150526, SMARTCyp incorrectly predicts the SoM as C3, however the docking results suggest that the conformation leading to metabolism of C3 is unfavorable (ranked 15th overall). The overall top-ranked pose (**d**) incorrectly predicts O14 as the true SoM, however this atom was not favorably ranked using SMARTCyp. The top-ranked bioactive pose (**e**) ranks the true SoM 4th overall and SMARTCyp ranks this atom 2nd overall. Combining these predictions leads to the true SoM, C20, being ranked 1st overall.

Although using a combination approach in the pseudo-apo ensemble improved performance over either SMARTCYP alone or docking alone, and all approaches tested on the crystal structure alone, we hypothesized that using Quantitative Structure-Activity Relationship (QSAR) modeling could improve the separation between active and inactive docking poses and further improve SoM prediction in the top-1, top-2, and top-3 positions.

4.3.6 Ranking CS data using QSAR

QSAR is a computational method that derives statistical relationships between sets of descriptors, typically ligand properties, and a set of values, typically the biological activities of the target ligands. We hypothesized that there were certain key ligand features, along with the spatial orientation of those features, which could distinguish between active and decoy docking poses, and that QSAR statistical modeling could be used to identify those features. By assigning a favorable score to active poses as compared to decoy poses, we aimed to train a model to preferentially select active ligand poses.

However, in addition to protein-ligand interactions, for CYP metabolism the reactivity of a chemical group is an additional critical factor to determine the potential SoM of a ligand. As pseudo-receptor QSAR programs, such as RAPTOR, do not directly incorporate this factor, we modified RAPTOR to include SMARTCyp scores as a descriptor in the modeling process. Using this modified QSAR approach, we were able to significantly improve SoM predictions (Table 4.2 - Vina+SMARTCyp+QSAR column).

Compared to SMARTCyp alone, docking alone, or the combined approach using SMARTCyp and docking (*CS*), re-ranking atoms using QSAR proved to be far superior. For example, the QSAR model based on the docking results from the pseudo-apo ensemble was able to predict the correct SoM in 88% and 96% within the top-1and top-2 positions, respectively.

For several compounds the QSAR approach drastically improved the ranking of the true SoM in comparison to the other methods tested (Figure 4.7). For instance, for etodolac (Figure 4.7a and 4.7b) none of the other methods tested accurately predicted the true SoM, C13, within the top-3 positions. However, using QSAR, the top ranked pose placed the true SoM within 4.0Å of the reactive oxygen. Notably, this pose was very poorly ranked using docking (10th overall). In other compounds, such as 17 alpha-ethinylestradiol (Figure 4.7c and 4.7d), the QSAR score offered incremental improvements within the top-3 ranked positions. In this compound, both docking and SMARTCyp were unable to rank the true-SoM within the top-3 positions. While the *CS* method improved the ranking to the top-2 position, QSAR ranked the true SoM at the top-1 position.

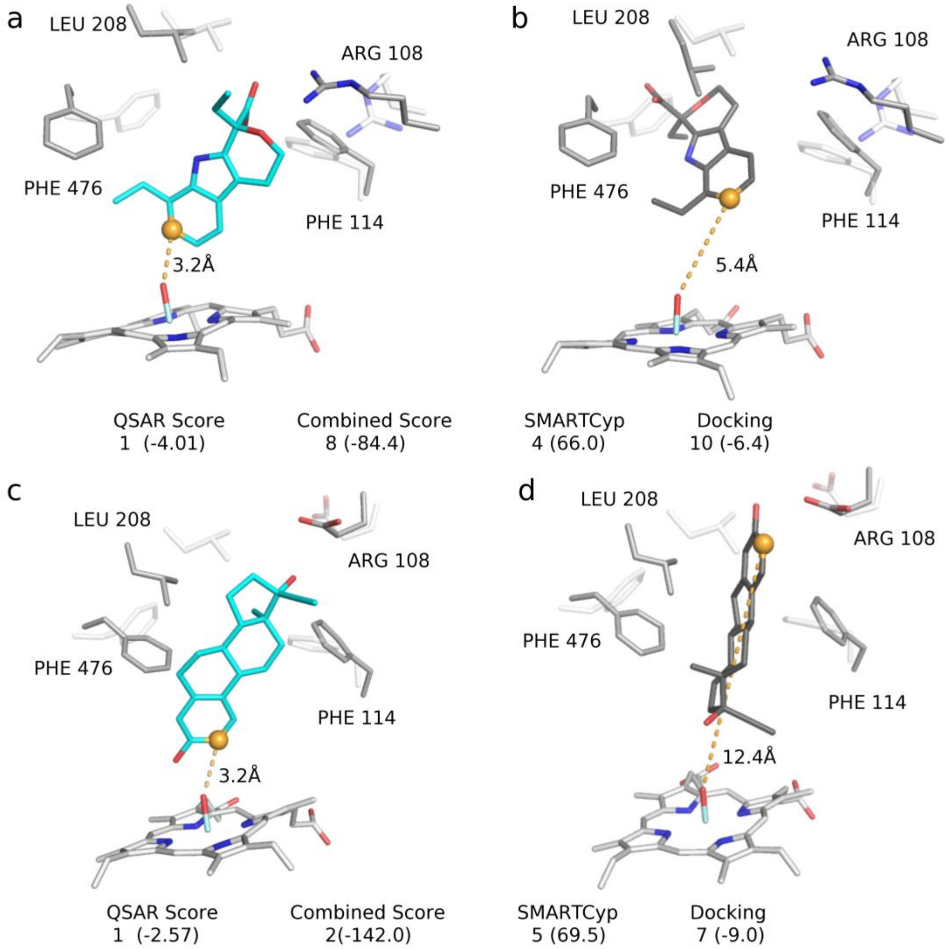


Figure 4.7: Examples of compounds in which the QSAR method improves SoM prediction over other tested methods. In the left column is the pose selected by the QSAR model and in the right column is the top-ranked docking pose, in both cases the true SoM has been shown in an orange sphere. The transparent white sticks represent the conformation of the crystal structure. In the case of Etodolac, the top ranked QSAR and top-ranked docking pose both have the true SoM oriented towards the reactive oxygen, but the QSAR pose selects the structure in which the SoM is within a reactive distance (**a** and **b**). For17 alpha-ethinylestradiol, the QSAR model selects a pose that is completely inverted (**c**) from the top-ranked docking pose (**d**). In both cases, the QSAR model places the known SoM in the top-1 predicted position. Notably, in all cases successful ligand docking requires a significant rearrangement of the binding site residues, as neither of these compounds could be successfully docked to the crystal structure.

One remaining limitation for the success of the QSAR model was the overall docking success. In other words, assuming that an active pose was sampled during the docking process, the QSAR model was nearly always able to identify the true SoM within the top-3 positions. In the crystal structure for instance, the QSAR model accurately predicted the SoM within the top-3 positions for all but one of the compounds for which an active docking pose was obtained. For the pseudo-apo ensemble, all compounds with an active docking pose were predicted within the top-2 ranked SoM.

Table 4.3 represents the QSAR results in isolation, i.e. only ligands for which active docking poses were found are considered. In this situation, 91%, 100% and 100% of the known SoMs are correctly predicted when the pseudo-apo ensemble was used for docking within the top-1, top-2 and top-3, respectively. These percentages are slightly lower when the crystal structure was used for docking, *i.e.*, 77%, 87%, and 98%, for the top-1, top-2, and top-3 positions, respectively.

While the QSAR model using the results from docking to the crystal structure was severely limited by the quality of the docking process, the pseudo-apo docking set was able to generate active poses for most ligands, allowing the subsequent QSAR model to predict the known SoM in the top-2 positions in 96% of all cases.

Table 4.3: QSAR SoM Rankings of ligands with an active docking pose.

	Fraction of Ligands with Rank			Total Number of Active Ligands
QSAR Model	Top 1	Top 2	Top 3	
X-ray	0.77	0.87	0.98	47
Pseudo-apo Ensemble	0.91	1.00	1.00	70

Over-fitting can be a concern in QSAR modelling, so the results for the test and training set of the QSAR models were compared (Table 4.4**).** Similar to Table 4.3, only those ligands for which an active docking pose was found are included in the comparisons. For the x-ray structure, there was little differences between the two sets; the training and test set had approximately the same fractions in the top-1, top-2, and top-3 positions. For the pseudo-apo ensemble, the prediction accuracy of the test set exceeded that of the training set, where the SoM of all ligands was correctly predicted in the top-1 position. This indicates that the chemical space of the test set was well-covered by the training set, and that the model has high predictive power for future compounds within the space modeled.

Additionally, as the RAPTOR QSAR package generates a pseudo-receptor model of the protein binding pocket around the ensemble of ligand poses, we visually compared the QSAR model with the members of the pseudo-apo structural ensemble, a representative example is shown in Figure 4.8. As shown, there is significant agreement between the protein structure and the RAPTOR model. Where the model predicts hydrophobic properties, the protein residues are mainly hydrophobic, such as LEU 366 and 361, and PHE 100, 114, and 476. Hydrophilic residues such as ARG 108, ASN 204, and ASP 293 are collocated with hydrophilic features of the RAPTOR pseudo-receptor. PHE 100 and 114 both appear to be able to engage in different types of interactions, as they are co-located with both hydrophobic and hydrophilic features, indicating that π stacking interactions might play an important role in the binding pocket.

Table 4.4: Comparison of QSAT Test and Training Sets.

	Fraction of Training Set Ligands with Rank			Total Number of Active Ligands
QSAR Model	Top 1	Top 2	Top 3	
X-ray, training	0.76	0.88	1.00	33
X-ray, test	0.79	0.86	0.93	14
Pseudo-apo, training	0.88	1.00	1.00	51
Pseudo-apo, test	1.00	1.00	1.00	19

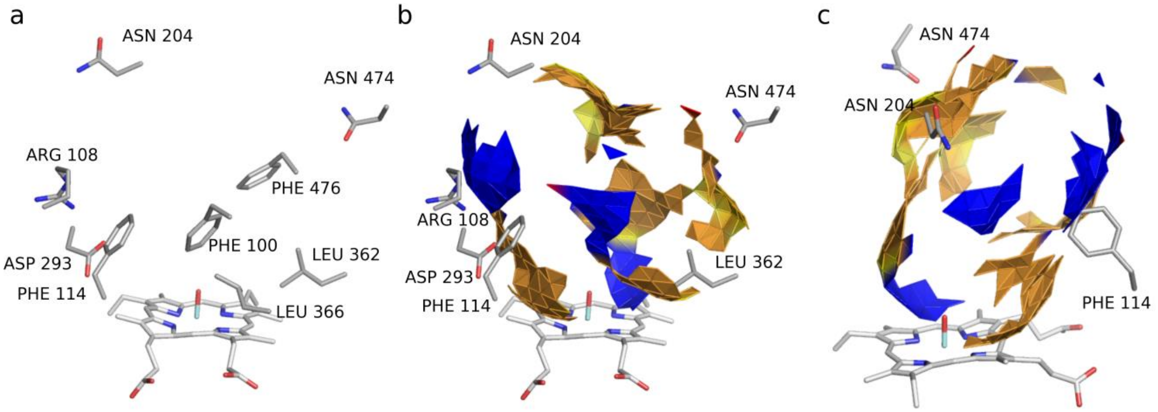


Figure 4.8: The QSAR model of the pseudo-apo data set. The binding site residues (**a**) and pseudo-receptor (**b**) with 90º rotation (**c**) generated by RAPTOR. The pseudo-receptor RAPTOR models are colored by property, with red representing hydrogen bond acceptors, blue hydrogen bond donors, and brown and yellow as hydrophobic regions.

4.4 CYP2C9 Conclusions

In this study, we compared the ability of ligand-based, structure-based, and combination-based approaches to predict the SoM in 73 diverse CYP2C9 substrates. Of all individual methods tested, docking was found to have the poorest performance. Whereas ensemble docking showed a significant improvement over docking to the crystal structure alone, at most 38% of the compounds were ranked in the top-1 position using docking. Using the SMARTCyp reactivity model alone, 42% of the compounds were accurately prediction in the top-1 position. By combining the docking scores and SMARTCyp scores prediction accuracy was improved in both ensembles, but not in the crystal structure. Ultimately, we found that the inclusion of QSAR into the combination approach resulted in significant improvement in prediction success and was the most effective and accurate SoM prediction method tested in this work.

In all systems tested, the QSAR model was able to accurately predict, within the top-3 positions, the SoM for nearly all ligands with an active pose. A key limitation to the success observed with QSAR was the ability of docking to provide active poses, in other words, poses in which the true SoM was within a reactive distance to the oxygen of the heme. Using a pseudo-apo ensemble, we were able to find an active docking pose for nearly all ligands tested. To set our results in perspective, a recent study of currently published methods found that accurate predictions in the top-2 positions range between 68-87%, on average, across various CYP isoforms[17]. In the same study, the highest prediction rate achieved for CYP2C9 was 87% in the top-2[17]. Using our approach we achieved an accurate prediction rate of 96% in the top-2 positions, albeit using a different, dataset.

Our promising results in CYP2C9 represent a step towards improved and highly accurate SoM predictions in CYP enzymes. While in the current study we tested substrates of CYP2C9, we believe that the proposed method will be of use in broader ligand datasets and also will be applicable to different CYP isozymes.

4.5 Extension of method to other CYP Isozymes

The general ligand preparation procedure for the other CYP isozymes was identical to the method used to prepare the CYP2C9 set. The size of the ligands sets for the CYPs is as follows: 1A2 (271 Ligands), 2A6 (105 Ligands), 2B6 (151 Ligands), 2C8 (141 Ligands), 2C19 (218 Ligands), 2-D6 (270 Ligands), 2E1 (145 Ligands), 3A4 (475 Ligands). The rest of the methods used to generate the SoM prediction models follow the same procedure as section 2 of this chapter with two major exceptions.

The first change is with regards to the ligands used to select the representative protein ensemble members. For 2C9, these ligands were manually selected. For the other data sets, this process was automated. The ligand sets were clustered based on similarity to select structurally diverse ligands. Generally, the size of the diverse selected was set to be approximately 10-20% of the total ligand set. This guided selection process was used for all remaining eight CYP isozymes. Similarly, for 2C9 the initial selection of the test and training sets was random. This random set was then manually curated for coverage of the chemical space of the ligand set. This process was also automated for the other ligand sets using the same clustering method used to select the ensemble selection set, only with slightly larger number of clusters. This process typically resulted in test sets of similar size to the diverse selection set, 15-20% of ligand compounds.

4.6 Results and Comparison for CYP Isozymes

As can be in Table 4.5, our models for all CYP isozymes produced highly reliable predictions of CYP SoMs. The Top-1 prediction rates range from 79% for 1A2 to 97% for 2B6. The Top-3 prediction rates range from 85.2% for 2-D6 to 100% for 2C8. The average prediction rate across all nine isozymes is 85% in the Top-1, 92% in the Top-2 and 93% in the Top-3. It is important to note that overall docking accuracy was comparable to the Top-3 percentage at 94%. Overall, and for each individual CYP isozyme, the same general trend noticed for CYP2C9 was observed: if an active docking pose for a ligand can be found, the QSAR model is generally able to place it in one of the top spots.

To investigate this issue, a secondary model was built for several of the CYP isozymes with the lowest docking accuracy, including CYP2-D6. For these secondary models, those ligands for each CYP isozyme which failed to find an active docking pose in the initial model were used as the screening ligands for a repeat of the protein ensemble member selection process. In general, most (over 90%) of the failed ligands successfully docked to an ensemble member at this stage. In addition, there was significant overlap between the original protein structure ensembles and the newly selected ensembles. The new structures were added to the original protein ensembles, and the rest of the method was repeated.

Table 4.5: SoM Prediction for Nine Cyp Isozymes

Cyp Isozyme	Number of Compounds	Succesful Docking %	Top-1	Top-2	Top-3
1A2	271	93	79	90	91
2A6	105	94	87	93	93
2B6	151	99	97	99	99
2C8	141	100	94	100	100
2C9	226	96	88	96	96
2C19	218	88	80	87	88
2D6	218	87	80	84	85
2E1	270	96	81	89	91
3A4	475	94	80	90	92

However, after the new models were completed, there was no significant change in overall docking accuracy or final prediction rates of the models. Comparing which ligands were successfully docked for each model, while most ligands docked successfully to both models, some ligands docked successfully to only one model, while some docked to neither. This last set was of particular interest, as many of these ligands docked successfully in the ensemble selection process. When the results of the ensemble selection process were further analyzed, we found a common characteristic: The ligands that successfully docked in the ensemble selection stage did so with low ranks. We therefore concluded that the docking scoring process was a weakness in our method, as it preferred non-active poses for some ligands.

Even with this issue with the docking process, our results compare favorably with the top methods in the field. Table 4.6 is a comparison of the Top-2 prediction percentages for our method, along with Xenosite[17], RS-Predictor[12], and SmartCYP[2], along with the random prediction rate. The major difference between CyPredict and the other methods is the use of structure-based information. CyPredict uses both ligand-based and structure-based methods, while the other methods are purely ligand-based, using quantum chemical and topological descriptors. Our method produced the highest successful prediction rate for seven of the nine tested CYP isozymes, with the exceptions being 2C19 and 2-D6. For CYP2C19, our rate was 87% as compared to 89% for Xenosite. For CYP2-D6, our rate was 84% compared to 89% for Xenosite and 86% for RS-Predictor. For the other isozymes, our method performed 3% to 15% better than the next best method, and our average prediction rate was 92% compared to 87% for Xenosite, 84% for RS-Predictor, and 82% for SMARTCyp.

Table 4.6: Comparison of Cyp SoM Prediction Methods

Method	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	Ave.
Lill	**90.0**	**93.0**	**98.7**	**100**	**96.0**	87.2	83.7	**89.0**	**90.0**	**92.0**
Xenosite[17]	87.1	85.7	83.4	88.7	86.7	**89.0**	**88.5**	83.5	87.6	87.0
RS-Predictor[12]	83.4	85.7	82.1	83.8	84.5	86.2	85.9	82.8	82.3	84.3
SMARTCyp[2]	80.0	86.0	77.0	83.0	84.0	86.0	83.0	82.0	78.0	82.1
Random	26.0	31.9	24.8	22.6	22.2	20.2	21.1	36.5	21.0	25.3

Another key factor in these comparisons is the validation method for the models. Xenosite used leave-one-out cross-validation, while RS-Predictor and SMARTCyp used 10-fold validation. These methods use multiple models with test sets of either one compound (leave-one-out) or 10% of compounds (10%). In comparison, our test sets range from approximately 50% to 90% of our ligands. These large test sets indicate that our models have retained significant predictive power while avoiding possible overfitting of the data.

4.7 Conclusions

With these last studies, we have shown that we are successfully able to extend our model to other CYP isozymes beyond 2C9. Our models compare favorably with the current best-performing CYP SoM prediction techniques, and in several cases significantly outperform them. In addition, we have identified a specific area of concern to focus on for improving our methods: increasing docking accuracy. Currently, this is the weakest portion of our method, as the pseudoreceptor modeling process is generally able to correctly select active ligand poses if one has been generated by the docking process. Beyond improving docking, any further improvements in the process will require more complex calculations, such as QM/MM methods, as the second largest source of error we found is in the SMARTCyp scoring process.

In addition, the completed CYP models are being made freely available to academic users for SoM prediction through a webserver using the Nanohub platform.[18] The server will be available at https://nanohub.org/tools/cypredict/.

List of References

1. Danielson ML, Desai PV, Mohutsky MA, Wrighton SA, Lill MA. Potentially increasing the metabolic stability of drug candidates via computational site of metabolism prediction by CYP2C9: The utility of incorporating protein flexibility via an ensemble of structures. Eur J Med Chem. 2011;46(9):3953-63.

2. Rydberg P, Gloriam DE, Zaretzki J, Breneman C, Olsen L. SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. ACS Med Chem Lett. 2010 2010/06/10;1(3):96-100.

3. Lill MA, Vedani A, Dobler M. Raptor: Combining Dual-Shell Representation, Induced-Fit Simulation, and Hydrophobicity Scoring in Receptor Modeling: Application toward the Simulation of Structurally Diverse Ligand Sets. J Med Chem. 2004 2004/12/01;47(25):6174-86.

4. Nebert, DW; Russell, DW. Clinical importance of the cytochromes P450. *The Lancet*. 360[9340], 1107-1182 (2002). 24. Sykes MJ, McKinnon RA, Miners JO. Prediction of Metabolism by Cytochrome P450 2C9: Alignment and Docking Studies of a Validated Database of Substrates. J Med Chem. 2008 2008/02/01;51(4):780-91.

5. Sykes MJ, McKinnon RA, Miners JO. Prediction of Metabolism by Cytochrome P450 2C9: Alignment and Docking Studies of a Validated Database of Substrates. J Med Chem. 2008 2008/02/01;51(4):780-91.

6. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol. 1999;285(4):1735-47.

7. Lill M, Danielson M. Computer-aided drug design platform using PyMOL. J Comput Aided Mol Des. 2011 2011/01/01;25(1):13-9.

8. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J Chem Theory Comput. 2008 2008/03/01;4(3):435-47.

9. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. J Comput Chem. 2005;26(16):1701-18.

10. Oda A, Yamaotsu N, Hirono S. New AMBER force field parameters of heme iron for cytochrome P450s determined by quantum chemical calculations of simplified models. J Comp Chem. 2005;26(8):818-26.

11. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A linear constraint solver for molecular simulations. J Comput Chem. 1997;18(12):1463-72.

12. Zaretzki J, Bergeron C, Rydberg P, Huang T-w, Bennett KP, Breneman CM. RS-Predictor: A New Tool for Predicting Sites of Cytochrome P450-Mediated Metabolism Applied to CYP 3A4. J Chem Inf Model. [doi: 10.1021/ci2000488]. 2011; 51(7):1667-89.

13. Rydberg P, Olsen L. Predicting Drug Metabolism by Cytochrome P450 2C9: Comparison with the 2D6 and 3A4 Isoforms. ChemMedChem. 2012;7(7):1202-9.

14. Ekroos M, Sjögren T. Structural basis for ligand promiscuity in cytochrome P450 3A4. Proc Natl Acad Sci USA. 2006 September 12, 2006; 103(37):13682-7.

15. Hritz J, de Ruiter A, Oostenbrink C. Impact of Plasticity and Flexibility on Docking Results for Cytochrome P450 2D6: A Combined Approach of Molecular Dynamics and Ligand Docking. J Med Chem. 2008; 51(23):7469-77.

16. Teague SJ. Implications of protein flexibility for drug discovery. Nat Rev Drug Discov. [10.1038/nrd1129]. 2003; 2(7):527-41.

17. Zaretzki J, Matlock M, Swamidass SJ. XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks. J Chem Inf Model. 2013 2013/12/23;53(12):3373-83.

18. Krishna M, Zentner L, *et al.* nanoHUB.org: Cloud-based Services for Nanoscale Modeling, Simulation, and Education. Nanotech. Rev. 2013; 2(1): 107–117

CHATER 5. FUTURE DIRECTIONS

5.1 Research Summary

The overall goal of my research was the application and development of the advancement of combined ligand-based and structure-based techniques, namely pseudoreceptor-based methods, with a focus on surface-based pseudoreceptors. While the goal of pseudoreceptor methods is to produce a protein-like structure to interact with ligands, there has been a lack of use of protein structural data in the guiding of the creation of the pseudoreceptors. In Chapter 2, analysis of the interaction surface between protein crystal structure and co-crystallized ligand for the refined set of the PDBbind database was presented. These surfaces represented the ideal pseudoreceptor, as they mapped the true interactions of protein and ligand, and the analysis showed that the majority of protein-ligand interactions can be mapped by a few of Gaussian-based descriptors that have parameters that fall into a small range of values. In Chapter 3, a means of tuning surface-based pseudoreceptors to accurately replicate protein binding pocket topology as from known binding ligands will be presented.

In Chapter 4, I will discuss the implementation of the refinement of our group's previous work on SoM prediction, which includes the use of a modified version of the RAPTOR pseudoreceptor package. The modification was the inclusion of reactivity scores from the SMARTCYP package as term in the RAPTOR scoring function. The motivation for the inclusion of RAPTOR was as a means of generating a model which could reliably select binding poses with the known SoM close to the heme of CYP. This method was implemented as a means to counteract the difficulties arising from the large number of poses generated by the ensemble docking process. The initial modeling was performed on CYP2C9, but was later extended to eight other CYP isozymes. In this final chapter, we will discuss several possible methods for continuing or improving upon the research discussed in this thesis.

5.2 Pseudoreceptor Method

To extend the work presented in Chapters 2 and 3, we have worked to implement a new pseudoreceptor-based QSAR package based on the RAPTOR package but with improvements based on the insights presented in this thesis. Significant progress has been made on this new computational tool, but it has not yet reached completion. In addition to improvements to the pseudoreceptor method, we have also chosen to move to Python from the C-based languages for the primary programming language. Python is well-suited for file and data management tasks, but lacks speed for intense computations. We have used *weave* to integrate fast C code for those portions of the code with large numbers of complex calculations. The method alterations are primarily focused on a Correlated Mutli-surface Model and an altered machine learning and scoring method.

5.2.1 Correlated Multi-surface (CMS) Model

After our analysis in Chapter 3, where we determined that a single iso-surface is unable to represent the flexible binding pocket of a protein for diverse ligands, we developed a CMS model. Our first thought was to simply use multiple independent iso-surfaces of varying iso-level values. Each iso-surface is generated from the ligand occupancy values via the Marching Cubes algorithm. After consideration though, we had a number of concerns with this process. With independent iso-surfaces, even regions where the surfaces are very close together, the algorithm could have assigned sets of radically different Gaussians to the surfaces, which is not realistic. If the shells are representing the same region and conformation of a protein, they should have identical physico-chemical (electrostatic, hydrophobic, hydrogen bond) properties. To address this, we decided to correlate the iso-surfaces if they are physically close to each other. To determine correlation, for each vertex of each iso-surface, the closest vertex (as determined by distance scaled by angle between the vertices) of every other iso-surface is found. If the closest vertex is within a certain cut-off, the vertices are then linked to each other, which is important when generating the initial Gaussians and in the genetic algorithm.

After generating and correlating our multiple iso-surfaces, we generate our initial Gaussian regions via the following process. First, we determine the total strength of every ligand atom's interactions with every shell that contains that atom. (A shell contains a ligand atom if the iso-level of that shell is lower than the occupancy value at that atom's coordinates.) This strength is determined by the same functions as used for the protein in the PLSIA algorithm presented in Chapter 2. From the ten vertices with the largest value for a property, one is randomly selected and a Gaussian is generated using random value parameters, with ranges determined from our previous work. This Gaussian is then mapped to the surrounding surface, and a Gaussian region is determined. If the surface point where the center of the Gaussian is correlated, this Gaussian is propagated to all correlated iso-surfaces. All members of the Gaussian region are then excluded, and the ten strongest non-excluded vertices are found and a new center chosen. This process repeats up to a maximum number of iterations for all properties of all shells, with correlated regions counting towards the maximum. In order to provide more diverse models, the maximum number of iterations is randomly determined.

5.2.2 Scoring and Machine Learning

The Gaussian-mapped shells are then passed to the machine learning algorithm for the creation of the final pseudoreceptor model. We have implemented the PyEvolve genetic algorithm package with customized functions. We have implemented correlated cross-over and mutation functions. In the cross-over function, the algorithm selects one Gaussian from one of the parent models and then swaps that Gaussian and all its correlated Gaussians with a set of correlated Gaussians of the appropriate physico-chemical property from the other parent. This cross-over is also restricted by a distance cut-off between the locations of centers of the Gaussians: only Gaussians located in the approximately the same location may be swapped. The mutation function also works amongst correlated Gaussians. The allowed mutations are addition, deletion, moving the center along a single edge of the iso-surface, and change of Gaussian parameter (radius, amplitude, and angle). These mutations are propagated to all correlated Gaussians, and correlated Gaussians are created or deleted if necessary.

After a certain number of steps, we export the top models from the genetic algorithm to a Monte Carlo optimization algorithm along with the initial ligand conformations from the alignment process. In the Monte Carlo process, the ligand conformations are allowed to translate as well as rotate. The best pose for each ligand conformation, as scored using the same scoring function as in the genetic algorithm, is returned to the pseudoreceptor program for all Monte Carlo models. The best scored pose for each ligand conformation from amongst all the Monte Carlo models is identified and selected. These poses are then used to generate a new pseudoreceptor model (new iso-surfaces and genetic models) until a set number of iterations of the full process have been completed, then the final predictions are generated.

These processes will be scored by computing the interaction between the fields from the Gaussian surfaces directly with the ligand atoms. Typically, scores are computed between the ligand atoms and the discrete grid points. This point-to-point pairwise scoring not only creates a large number of descriptors which can lead to overfitting; it does not replicate the surfaces found in actual protein-ligand interactions. With our process, instead of each grid point being independent, regions of the iso-surface are linked by Gaussian functions where the properties of the entire region are determined by the four 2-D Gaussian parameters. We are currently scoring the Gaussians with a simple modified piece-wise linear pairwise (PLP) scoring function combined with an electrostatic term. The electrostatics are computed by computing the pairwise Coulombic interaction between the ligand atoms and the iso-surface vertices of electrostatic Gaussians. Hydrophobic, steric, and hydrogen bond terms are calculated using a modified PLP function (Gehlhaar, 1995). The PLP scoring function follows the general form shown in Figure 5.1, and is calculated between ligand atoms and the iso-surface vertices appropriate Gaussian regions, except the steric term is calculated with the full iso-surface instead of Gaussian regions. The individual interactions are then scaled according the Gaussian value of the vertex.

5.3 CYP SoM Prediction

As illustrated in Chapter 4, we have achieved significant success with our algorithm for the prediction of SoMs for a number of CYP isozymes, but we also feel there are potential avenues to improve and extend the method. First, as mentioned previously, one major issue is docking accuracy, which is typically less than 100% for the CYP isozymes. Second, while we have tested the performance on a single CYP isozyme at a time, we have not explored trying to predict SoM's against multiple CYP isozymes simultaneously.

5.3.1 Docking Accuracy

As previously mentioned, with our method, if we can obtain an active docking pose, the pseudoreceptor model can generally identify that pose. Therefore, the major source of error for our models is in the docking process. We found that increasing the number of protein ensemble members was not successful in remedying this issue. This is supported by our studies of CYP2C9, where we found insignificant increase in model accuracy when using more than six protein structure models. The other CYP isozymes seem to follow this trend, adding structures above a certain minimum does not significantly improve docking accuracy.

During our analysis of our results, we found the ligands that were not always successfully docked often had active poses that were poorly ranked. Currently, we only take the top ten ligand poses into consideration, and the active poses for these ligands are generally ranked close to ten if successfully docked at all. Therefore, to increase docking accuracy, one possible solution is to increase the number of generated docking poses. This is not an ideal solution, as this increases the number of poor poses for those ligands for which we can find active poses in the top ten. This can eventually cause difficulties for the pseudoreceptor modeling process. Another possible solution is iterating the docking process. As we know from our previous studies, we can occasionally find active poses from the difficult ligands. Therefore, if we run the docking process multiple times, an active pose may be generated. As we would still only generate the top ten poses for each iteration, the ratio of active to decoy poses should remain relatively constant, which should theoretically be favorable to purely increasing the number of accepted poses. That is, while both processes could decrease QSAR accuracy due to an increased number of docking poses, the iterative process should have a better ratio of active to decoy poses, so long as the docking scoring process is better than random and an equivalent total number of poses are generated.. Extensive docking studies are needed to determine which method is preferred, as it is dependent on both how likely we are to find an active pose at a given rank, and how this changes when the total number of ranks is increased. Another alternative is an additional docking processing step where the poses are scored with a more accurate function, such as MM/PBSA or MM/GBSA. While using more sophisticated scoring function would increase the computational cost of docking, it might be possible to maintain or reduce the number of necessary poses.

5.3.2 CYP Selectivity

The percentage of approved drugs that the CYP isozymes, studied in this thesis, metabolize was discussed in Chapter 4. The sum of these percentages is well above 100%, which is indicative of one of the issues of metabolism: multiple metabolic pathways. In our studies, we have only worked with known ligands for each CYP isozyme. While being able to reliably predict where each isozyme will metabolize a ligand, it is also important to be able to predict which isozymes will metabolize a given ligand.

In order to assess the ability of our method to address this problem, instead of using a tailored ligand set for each isozyme, we will repeat our studies with a combined ligand set formed from the individual isozyme sets. For each isozyme, those ligands that are not known to be metabolized by that isozyme will have all their docking poses classified as decoy poses. The selectivity of our method for each isozyme will then be determined by analyzing the scores of the known ligands, most likely using a method such as a receiver operating characteristic curve. This analysis would give a score of one if our method scores all known metabolites ahead of all known inactive compounds, and gives a score of zero for the reverse scenario.

APPENDIX

APPENDIX

This appendix contains supplementary tables S1-S11 for Chapter 3.

Table A1: Protein systems and corresponding pdb codes

Protein System	**PDB Entries**
HIV-PR	1a94
	1aaq
	1g2k
	1g35
	1gnm
	1gnn
	1gno
	1hbv
ERα	1gwq
	1gwr
	1x7e
	2p15
	2q70
	2qe4
	2qgt
CDK - 20	1b38
	1h00
	1h0v
	1h0w
	1h07
	1ke5
	1pxm
	1pnx
	1pxp
	1q8t
	1q8u
	1q8w
	1rej
	1stc
	1vyz
	1y91
	1yds
	1ydt
	2uzn
	2uzo

Protein System	PDB Entries
CDK - 10	1q8u 1q8w 1rej 1stc 1vyz 1y91 1yds 1ydt 2uzn 2uzo
CDK - 5	1y91 1yds 1ydt 2uzn 2uzo

Table A2: Occupancy Distribution of Estrogen Receptor

c-value	Fraction of surface points with target occupancy																				
1.4	0.001	0.072	0.183	0.167	0.153	0.126	0.096	0.073	0.044	0.026	0.017	0.015	0.008	0.009	0.006	0.002	0.001	0.002	0.000	0.000	0.000	0.000
2	0.000	0.012	0.052	0.083	0.102	0.122	0.128	0.112	0.106	0.081	0.074	0.045	0.029	0.018	0.016	0.008	0.005	0.005	0.002	0.000	0.000	0.000
2.5	0.000	0.002	0.028	0.033	0.041	0.067	0.091	0.103	0.118	0.111	0.103	0.090	0.084	0.050	0.027	0.020	0.014	0.012	0.003	0.002	0.000	0.000
3	0.000	0.000	0.016	0.030	0.016	0.017	0.044	0.068	0.081	0.104	0.122	0.115	0.111	0.095	0.077	0.045	0.027	0.017	0.009	0.005	0.000	0.000
Target	0	0.025	0.075	0.125	0.175	0.225	0.275	0.325	0.375	0.425	0.475	0.525	0.575	0.625	0.675	0.725	0.775	0.825	0.875	0.925	0.975	1.000

Table A3: Cumulative Occupancy of Estrogen Receptor

c-value	Cumulative Occupancy Fraction																				
1.4	1.000	0.999	0.927	0.744	0.577	0.424	0.298	0.202	0.129	0.085	0.059	0.042	0.028	0.019	0.011	0.005	0.003	0.002	0.000	0.000	0.000	0.000
2	1.000	1.000	0.988	0.936	0.854	0.751	0.630	0.502	0.390	0.284	0.203	0.128	0.084	0.055	0.036	0.021	0.013	0.008	0.003	0.000	0.000	0.000
2.5	1.000	1.000	0.998	0.969	0.936	0.895	0.827	0.737	0.634	0.516	0.406	0.302	0.212	0.128	0.078	0.051	0.031	0.017	0.005	0.002	0.000	0.000
3	1.000	1.000	1.000	0.984	0.954	0.939	0.922	0.877	0.809	0.729	0.624	0.502	0.387	0.275	0.180	0.103	0.058	0.031	0.014	0.005	0.000	0.000
Target	0.000	0.025	0.075	0.125	0.175	0.225	0.275	0.325	0.375	0.425	0.475	0.525	0.575	0.625	0.675	0.725	0.775	0.825	0.875	0.925	0.975	1.000

Table A4: Occupancy Distribution of HIV-PR

c-value	Fraction of surface points with target occupancy																				
1.4	0.049	0.146	0.151	0.147	0.137	0.114	0.088	0.064	0.044	0.026	0.016	0.009	0.005	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.011	0.095	0.090	0.083	0.089	0.102	0.108	0.105	0.089	0.072	0.059	0.042	0.028	0.017	0.008	0.003	0.000	0.000	0.000	0.000	0.000	0.000
2.5	0.001	0.059	0.063	0.063	0.059	0.066	0.077	0.092	0.098	0.096	0.092	0.076	0.064	0.045	0.028	0.015	0.005	0.001	0.000	0.000	0.000	0.000
3	0.000	0.029	0.045	0.048	0.048	0.048	0.051	0.065	0.076	0.091	0.096	0.100	0.094	0.081	0.060	0.043	0.021	0.006	0.000	0.000	0.000	0.000
Target	0.000	0.025	0.075	0.125	0.175	0.225	0.275	0.325	0.375	0.425	0.475	0.525	0.575	0.625	0.675	0.725	0.775	0.825	0.875	0.925	0.975	1.000

Table A5: Cumulative Occupancy of HIV-PR

c-value	Cumulative Occupancy Fraction																				
1.4	1.000	0.951	0.805	0.654	0.507	0.369	0.255	0.167	0.103	0.059	0.032	0.016	0.007	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	1.000	0.989	0.894	0.804	0.721	0.632	0.530	0.422	0.317	0.228	0.156	0.097	0.055	0.027	0.010	0.003	0.000	0.000	0.000	0.000	0.000	0.000
2.5	1.000	0.999	0.941	0.878	0.815	0.755	0.689	0.612	0.520	0.422	0.326	0.234	0.158	0.094	0.049	0.021	0.005	0.001	0.000	0.000	0.000	0.000
3	1.000	1.000	0.971	0.927	0.879	0.831	0.783	0.732	0.668	0.592	0.500	0.404	0.305	0.211	0.130	0.071	0.028	0.006	0.000	0.000	0.000	0.000
Target	0.000	0.025	0.075	0.125	0.175	0.225	0.275	0.325	0.375	0.425	0.475	0.525	0.575	0.625	0.675	0.725	0.775	0.825	0.875	0.925	0.975	1.000

Table A6: Occupancy Distribution of CDK-20

c-value	Fraction of surface points with target occupancy																				
1.4	0.014	0.182	0.221	0.150	0.105	0.077	0.062	0.051	0.043	0.037	0.026	0.018	0.009	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.003	0.095	0.143	0.136	0.125	0.097	0.082	0.069	0.058	0.048	0.040	0.038	0.030	0.023	0.010	0.003	0.000	0.000	0.000	0.000	0.000	0.000
2.5	0.000	0.058	0.095	0.097	0.108	0.105	0.096	0.080	0.073	0.061	0.052	0.044	0.042	0.038	0.029	0.017	0.005	0.000	0.000	0.000	0.000	0.000
3	0.000	0.033	0.070	0.063	0.077	0.096	0.093	0.094	0.079	0.076	0.066	0.057	0.047	0.045	0.044	0.037	0.020	0.003	0.000	0.000	0.000	0.000
Target	0	0.025	0.075	0.125	0.175	0.225	0.275	0.325	0.375	0.425	0.475	0.525	0.575	0.625	0.675	0.725	0.775	0.825	0.875	0.925	0.975	1.000

Table A7: Cumulative Occupancy of CDK-20

c-value	Cumulative Occupancy Fraction																				
1.4	1.000	0.986	0.804	0.583	0.433	0.328	0.251	0.189	0.138	0.095	0.058	0.031	0.013	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	1.000	0.997	0.902	0.759	0.623	0.499	0.402	0.320	0.251	0.193	0.146	0.105	0.067	0.037	0.014	0.004	0.000	0.000	0.000	0.000	0.000	0.000
2.5	1.000	1.000	0.942	0.846	0.750	0.642	0.537	0.440	0.361	0.288	0.227	0.175	0.131	0.089	0.051	0.022	0.005	0.000	0.000	0.000	0.000	0.000
3	1.000	1.000	0.967	0.897	0.834	0.757	0.662	0.569	0.475	0.395	0.319	0.253	0.196	0.149	0.104	0.060	0.023	0.003	0.000	0.000	0.000	0.000
Target	0.000	0.025	0.075	0.125	0.175	0.225	0.275	0.325	0.375	0.425	0.475	0.525	0.575	0.625	0.675	0.725	0.775	0.825	0.875	0.925	0.975	1.000

Table A8: Occupancy Distribution of CDK-10

c-value	Fraction of surface points with target occupancy																					
1.4	0.043	0.160	0.188	0.153	0.102	0.072	0.058	0.045	0.043	0.037	0.032	0.026	0.019	0.010	0.007	0.003	0.002	0.000	0.000	0.000	0.000	0.000
2	0.022	0.088	0.112	0.133	0.116	0.097	0.076	0.060	0.052	0.048	0.042	0.041	0.035	0.030	0.023	0.013	0.008	0.003	0.001	0.000	0.000	0.000
2.5	0.012	0.055	0.072	0.088	0.106	0.102	0.087	0.076	0.064	0.053	0.052	0.049	0.043	0.043	0.035	0.030	0.019	0.010	0.004	0.000	0.000	0.000
3	0.008	0.034	0.050	0.060	0.072	0.084	0.095	0.082	0.074	0.067	0.058	0.054	0.053	0.050	0.047	0.042	0.035	0.022	0.010	0.002	0.000	0.000
Target	0.000	0.025	0.075	0.125	0.175	0.225	0.275	0.325	0.375	0.425	0.475	0.525	0.575	0.625	0.675	0.725	0.775	0.825	0.875	0.925	0.975	1.000

Table A9: Cumulative Occupancy of CDK-10

c-value	Cumulative Occupancy Fraction																				
1.4	1.000	0.957	0.797	0.608	0.456	0.354	0.282	0.223	0.178	0.135	0.098	0.066	0.040	0.021	0.012	0.005	0.002	0.000	0.000	0.000	0.000	0.000
2	1.000	0.978	0.890	0.778	0.645	0.529	0.432	0.356	0.296	0.244	0.196	0.154	0.112	0.077	0.047	0.024	0.012	0.004	0.001	0.000	0.000	0.000
2.5	1.000	0.988	0.933	0.860	0.772	0.666	0.564	0.477	0.401	0.338	0.284	0.232	0.183	0.140	0.097	0.063	0.033	0.014	0.004	0.000	0.000	0.000
3	1.000	0.992	0.958	0.908	0.848	0.776	0.692	0.597	0.515	0.441	0.374	0.315	0.261	0.208	0.158	0.112	0.069	0.034	0.012	0.002	0.000	0.000
Target	0.000	0.025	0.075	0.125	0.175	0.225	0.275	0.325	0.375	0.425	0.475	0.525	0.575	0.625	0.675	0.725	0.775	0.825	0.875	0.925	0.975	1.000

Table A10: Occupancy Distribution of CDK-5

c-value	Fraction of surface points with target occupancy																					
1.4	0.080	0.168	0.154	0.126	0.098	0.073	0.056	0.050	0.045	0.036	0.034	0.032	0.022	0.012	0.007	0.004	0.002	0.000	0.000	0.000	0.000	0.000
2	0.040	0.102	0.100	0.100	0.100	0.093	0.079	0.068	0.056	0.050	0.045	0.042	0.036	0.032	0.027	0.017	0.008	0.004	0.001	0.000	0.000	0.000
2.5	0.024	0.068	0.065	0.075	0.081	0.081	0.086	0.080	0.074	0.066	0.055	0.049	0.046	0.040	0.034	0.034	0.025	0.012	0.005	0.001	0.000	0.000
3	0.013	0.045	0.049	0.051	0.060	0.065	0.071	0.078	0.080	0.077	0.074	0.063	0.054	0.049	0.043	0.041	0.042	0.030	0.012	0.003	0.000	0.000
Occupancy	0.000	0.025	0.075	0.125	0.175	0.225	0.275	0.325	0.375	0.425	0.475	0.525	0.575	0.625	0.675	0.725	0.775	0.825	0.875	0.925	0.975	1.000

Table A11: Cumulative Occupancy of CDK-5

c-value	Cumulative Occupancy Fraction																					
1.4	1.000	0.920	0.751	0.597	0.471	0.373	0.300	0.244	0.194	0.149	0.114	0.080	0.048	0.026	0.014	0.007	0.003	0.000	0.000	0.000	0.000	0.000
2	1.000	0.960	0.858	0.758	0.658	0.558	0.465	0.387	0.318	0.262	0.212	0.167	0.126	0.090	0.058	0.031	0.014	0.005	0.001	0.000	0.000	0.000
2.5	1.000	0.976	0.908	0.843	0.768	0.687	0.606	0.520	0.441	0.367	0.301	0.246	0.196	0.151	0.111	0.077	0.043	0.018	0.006	0.001	0.000	0.000
3	1.000	0.987	0.941	0.892	0.842	0.782	0.717	0.646	0.568	0.488	0.410	0.336	0.273	0.219	0.171	0.128	0.087	0.045	0.015	0.003	0.000	0.000
Target	0.000	0.025	0.075	0.125	0.175	0.225	0.275	0.325	0.375	0.425	0.475	0.525	0.575	0.625	0.675	0.725	0.775	0.825	0.875	0.925	0.975	1.000

VITA

VITA

Gregory Lee Wilson was born on July 14th, 1984 in Newton, Iowa, United States. He is the son of Gerald and Constance Wilson.

Gregory graduated from Penn High School in Mishawka, Indiana as valedictorian in 2003. He then attended Purdue University, earning a Bachelor of Science in Chemical Engineering in 2007.

After earning his undergraduate degree, Gregory continued his education at Purdue University, entering the Ph. D. program of the Medicinal Chemistry and Molecular Pharmacology Department. There, under the supervision of Dr. Markus Lill, he worked on the development of application of novel pseudoreceptor methods. After completing his Ph.D. studies, Gregory intends to continue working in the field of computer-aided drug design.

.

PUBLICATIONS

PUBLICATIONS

Wilson, GL.; Lill, MA. Integrating structure-based and ligand-based approaches for computational drug design, *Future Medicinal Chemistry*, **2011**, *3*, 735-770.

Wilson, GL.; Lill, MA. Towards a realistic representation in surface-based pseudoreceptor modelling: a PDB-wide analysis of binding pockets, *Molecular Informatics*, **2012,** *31*, 259-271

Gibbs et al., Compounds and methods for use in treating neoplasia and cancer, WO 2013016531 A3, **2013**

Wilson, GL.; Lill, MA. Integrating structure-based and ligand-based approaches for computer-aided drug design. In MA Lill (Ed.), *In silico drug discovery and delivery*, **2013**, 190-202

Kingsley, LJ.; Wilson, GL.; Essex, ME.; Lill, MA. Combining Structure- and Ligand-Based Approaches to Improve Site of Metabolism Prediction in CYP2C9 Substrates. *Pharm. Res.*, **2015**, *32*, 986-1001.

Wilson, GL.; Lill, MA. Surface-based pseudoreceptor modeling: Optimizing surface representations of binding sites using experimental protein-ligand structure data. Manuscript written