Applications and Improvements in the Molecular Modeling of Protein and Ligand Interactions

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Jason Bret Harris
August 2014

Copyright © 2014 by Jason Bret Harris
All rights reserved.

DEDICATION

I dedicate this work to my mother Susan and siblings Joshua, Sophia, and Shayne.

ACKNOWLEDGEMENTS

I would like to acknowledge and thank my committee members: Jeremy Smith, Gary Sayler, Elias Fernandez, and my advisor Jerome Baudry for their support and guidance. I also wish to recognize other mentors: Elizabeth Howell, Robert Hinde, Alice Layton, Melanie Eldridge, and Valerie Berthelier for their contributions to my research training. Finally, I would like to thank my peer fellowship members from the NSF training program (SCALE-IT) for their support and also the program manager Harry Richards and program director Cynthia Peterson for their influence in my graduate student training.

ABSTRACT

Understanding protein and ligand interactions is fundamental to treat disease and avoid toxicity in biological organisms. Molecular modeling is a helpful but imperfect tool used in computer-aided toxicology and drug discovery. In this work, molecular docking and structural informatics have been integrated with other modeling methods and physical experiments to better understand and improve predictions for protein and ligand interactions. Results presented as part of this research include:

1.) an application of single-protein docking for an intermediate state structure, specifically, modeling an intermediate state structure of alpha-1-antitrypsin and using the resulting model to virtually screen for chemical inhibitors that can treat alpha-1-antitrypsin deficiency,

2.) an application of multi-protein docking and metabolism prediction, specifically, modeling the cytochrome P450 metabolism and estrogen receptor activity of an environmental pollutant (PCB-30), and

3.) providing evidence to support the inclusion of anion-pi interactions in molecular modeling by demonstrating the biological roles of anion-pi interactions in stabilizing protein and protein-ligand structures.

This work has direct applications for mitigating disease and toxicity, but it also demonstrates useful ways of integrating computational and experimental data to improve upon modeling protein and ligand interactions.

TABLE OF CONTENTS

INTRODUCTION 1
Prologue 1
Scope of Dissertation 2
Introduction to Virtual Docking 2
Why Use Predictive Docking? 3
Introduction to Chapter 1: Modeling of an Intermediate State Structure 4
Introduction to Chapter 2: Multi-Protein Modeling and Metabolism Prediction 5
Introduction to Chapter 3: Anion-pi Interactions in Molecular Modeling 5
References 6
CHAPTER 1.1. DISCOVERY OF A POTENT INHITOR OF Z-ALPHA1 ANTITRYPSIN POLYMERIZATION 8
Abstract 9
Introduction 9
Materials and Methods 11
General Materials and Methods 11
Preparation of the bPEG-peptide 11
Preparation of Working Compound Plates 11
Set up of the Microplate Screening Assay 12
Z-α1AT Polymerization Inhibition Assay 12
Determining IC_{50} Values of Inhibitors 12
Z-α1AT Polymerization Sedimentation Assay 13
Preparing Protein Structures for Homology Modeling and Docking Simulations 13
Homology Modeling Procedure 13
Docking Simulation Procedure 13
Virtual Screening the NCI Diversity Database 14
Results 14
Principal Characteristics of the Z-α1AT Polymerization Inhibitor Screening Assay 14
S-(4-nitrobenzyl)-6-thioguanosine Identified as Inhibitor of Z-α1AT Polymerization 16
Validation of the Action of S-(4-nitrobenzyl)-6-thioguanosine 17
Structural Modeling of α1AT and the M* Intermediate 18
Analysis of α1AT Structures and their Potential Binding Sites 20
S-(4-nitrobenzyl)-6-thioguanosine Binds at the RCL Insertion Site or on the Edge of β-sheet A 21
Residue Interactions with S-(4-nitrobenzyl)-6-thioguanosine 21
Virtual Screening 24
Discussion 24
Acknowledgement 27
Funding Sources 27
Abbreviations 27

Supporting Information 27
References 32
CHAPTER 2.1. A COMPUTATIONAL APPROACH PREDICTING CYP450 METABOLSIM AND ESTROGENIC ACTIVITY OF AN ENDOCRINE DISRPUTING COMPOUND (PCB-30) 36
Abstract 37
Introduction 37
Materials and Methods 40
Ligand Preparation 40
P450 Docking 41
Estrogen Receptor Docking 41
SMARTCyp 42
Yeast Bioassays 42
Microsomal Reaction Mixture 43
Gas Chromatography/Mass Spectrometry 43
Results 44
In Silico SMARTCyp Predictions 44
In Silico P450 Docking Predictions 44
In Silico Estrogen Receptor Docking Predictions 44
In Vitro P450 Exposure and Bioassays 47
Gas Chromatography/Mass Spectrometry 48
Discussion 50
Conclusion 51
Acknowledgment 52
Abbreviations 52
Supplemental Figures 52
References 58
CHAPTER 3.1. A SURVERY OF ASPARTATE-PHENYLALANINE AND GLUTAMATE-PHENYLALANINE INTERACTIONS IN THE PROTEIN DATA BANK: SEARCHING FOR ANION- π PAIRS 64
Abstract 65
Introduction 65
Materials and Methods 67
Results 70
Discussion 82
Acknowledgment 84
Abbreviations 84
Supplemental Figures 85
References 89
CHAPTER 3.2. STAAR: STATISTICAL ANALYSIS OF AROMATIC RINGS 97
Abstract 98
Introduction 98
Methodology 99
Process 99
Web Service 102
Code 103

Results 103
Conclusion 105
Acknowledgment 105
References 106
CHAPTER 3.3. ANION-PI GEOMETRIES BETWEEN PROTEIN AND LIGAND STRUCTURES 108
Abstract 109
Introduction 109
Methods 111
General Methodology 111
Algorithm for Identifying 6-Carbon Aromatic Rings in Ligand Structures 111
STAAR Program 111
Results 111
Discussion 114
Acknowledgment 115
References 116
CONCLUSION 117
VITA 119

LIST OF TABLES

Table 1.1.1. Docking results from M-α1AT, Z-α1AT, and M* model with S-(4-nitrobenzyl)-6-thioguanosine and the 79 other small molecules. 26
Table 1.1.2. Residues interacting with S-(4-nitrobenzyl)-6-thioguanosine in top scoring binding sites. 26
Table 2.1.1. SMARTCyp: Reactive Atom Sites. 45
Table 2.1.2. P450 Docking (2D6 and 3A4): Atom Accessibility 46
Table 2.1.3. Docking (hER-α): Binding Predictions for PCB-30 and Metabolites 46
Table 3.1.1. Correlation coefficients between calculated interaction energies with various quantum mechanical treatments and CHARMM22. 78
Table 3.2.1. GAMESS input parameters.. 101
Table 3.3.1. Summary of Anion-pi Statistics.....112

LIST OF FIGURES

Figure 1.1.1 Kinetic diagram of bPEG-peptide binding to α1AT 15
Figure 1.1.2. Pattern of inhibition resulting from the screening of 80 unknown LOPAC compounds... 16
Figure 1.1.3. S-(4-nitrobenzyl)-6-thioguanosine inhibits bPEG-peptide binding to Z-α1AT..... 17
Figure 1.1.4. Effect of S-(4-nitrobenzyl)-6-thioguanosine on Z-α1AT polymerization. 18
Figure 1.1.5. The three models of α1AT protein..... 19
Figure 1.1.6. The fragments of structures 1QLP (green) and 3T1P (red) used to homology model the M* intermediate state of α1AT..... 20
Figure 1.1.7. Binding Sites for S-(4-nitrobenzyl)-6-thioguanosine.....22
Figure 1.1.8. Two crystal structures for α1AT are superimposed and represented in ribbon diagram.....23
Figure 1.1.9. Virtual Screening.....23
Figure 1.1.S1. 2-D contour and interaction map generated in MOE for S-(4-Nitrobenzyl)-6-thioguanosine at SITE1 in the M* intermediate state structure 28
Figure 1.1.S2. 2-D contour and interaction map generated in MOE for for S-(4-Nitrobenzyl)-6-thioguanosine at SITE2 in the M* intermediate state structure 29
Figure 1.1.S3. 2-D contour and interaction map generated in MOE for S-(4-Nitrobenzyl)-6-thioguanosine at SITE5 in the M* intermediate state structure 30
Figure 1.1.S4. 2-D contour and interaction map generated in MOE for S-(4-Nitrobenzyl)-6-thioguanosine at SITE6 in the 3CWM structure 31
Figure 2.1.1. Computational and Experimental Approaches.....39
Figure 2.1.2. Docking (hER-α): Atom Interactions. 47
Figure 2.1.3. Bioassay: Response to Standards (PCB-30 and 4-hydroxy-PCB-30). 48
Figure 2.1.4. Bioassay: Response to PCB-30 Metabolites (MRMs).....49
Figure 2.1.5. GC/MS: Characterization of PCB-30 Metabolites (MRMs).....50
Figure 2.1.S1. Computational Prioritization of the DUD-ER Database. 52
Figure 2.1.S2. Docking Pose for PCB-30 Atoms C.4, C.3, C.2 in CYP3A4. 53
Figure 2.1.S3. Docking Pose for PCB-30 Atoms C.3' in CYP3A4.....54
Figure 2.1.S4. Docking Pose for PCB-30 Atoms C.4 and C.3 in CYP2D6.....55
Figure 2.1.S5. Docking Pose for PCB-30 Atom C.2 in CYP2D6. 56
Figure 2.1.S6. Docking Pose for PCB-30 Atom C.3' in CYP2D6. 57
Figure 3.1.1. A flowchart of the calculations..... 68
Figure 3.1.2. Benzene-formate pairs. 70
Figure 3.1.3. Correlation between Hartree-Fock (HF) and Kitaura-Morokuma energies. 71
Figure 3.1.4. Figure 4. A histogram analysis of the interaction energies as a function of angle..... 73
Figure 3.1.5. Distance relationships of the BF pairs. 74
Figure 3.1.6. Structural analysis of 8260 BF pairs possessing energies from -8 to -2 kcal/mol.. 76

Figure 3.1.7. Depth of the top 100 BF protein pairs calculated by the PSAIAA program. 77
Figure 3.1.8. Interesting arrangements of anion-π pairs. 79
Figure 3.1.9. vs. CHARMM22 interaction energies, for functional groups.....81
Figure 3.1.S1. Plot of total interaction energies vs. “mix” term values. 85
Figure 3.1.S2. The fractional occurrence of BF pairs as a function of θ.....86
Figure 3.1.S3. Plot of distance differences vs. total interaction energy.....87
Figure 3.1.S4. Depth of BF pairs in the protein structures. 88
Figure 3.2.1. Flowchart of the STAAR program.....100
Figure 3.2.2. Web implementation of STAAR.....102
Figure 3.2.3. Angle and distance distributions for STAAR-identified anion-p pairs in the PDB. 104
Figure 3.2.4. Distribution of KM energies for STAAR-identified anion-π pairs in the PDB. 105
Figure 3.3.1. Principles of Anion-pi Interaction between benzene and formate.110
Figure 3.3.2. Anion-pi Pairs Involving Ligands with 1 to 6 Rings. 113
Figure 3.3.3. Anion-pi Pair Frequencies by Angle (10° increments). 113
Figure 3.3.4. Anion-pi Distance and Angle Distribution. 114

INTRODUCTION

Prologue

The structures of any given biological system are organized at different scales, and within each scale are various interactions that control how higher-order structure forms. In the same way, the structures from any given scale can be rationalized by the interactions of its substructures. The interactions between organisms and their environment represent the highest-order of biological organization. Cellular structures and interactions represent a middle area of organization. At the lowest scale are atomistic-level interactions between organic and inorganic molecular structures. One of the grand challenges in biology is to understand higher-order life processes from the organization of these molecular structures[1].

A typical eukaryotic cell contains billions of molecules. There are different macromolecules such as DNA, RNA, proteins, polysaccharides as well as many bioactive small chemical structures that interact with macromolecules. Proteins represent the largest and arguably the most functional group of macromolecules with an estimated 7.9×10^9 molecules per cell.[2] These structures are mostly studied in the context of interactions that lead to disease and toxicity. In order to approach a more complete molecular perspective of biology, new methods to efficiently study challenging molecular structures such as proteins and their interactions need to be developed.

All structures, including proteins, can exist in functionally different conformational states based on their inter- and intra-molecular interactions. Molecular and biophysical experiments can account for the net effects of structural interactions in a molecule or for a system of molecules, but they are unable to physically isolate all molecular structures or practically manage the evaluation for all of available structures in biology. Computational techniques that simulate molecular interactions can assist in expanding coverage for structures that are unresolved through experiments, but the sheer number of molecules and potential conformational states associated with macromolecular structures presents a sampling challenge for the reproducibility and reliability of simulations. In order to efficiently sample conformational space for even a single macromolecule, simulations rely on approximations of covalent and noncovalent molecular interactions and geometrical structural features that are believed to be relevant for capturing the energetic and thermodynamic properties of molecules. Continued research is needed to improve these approximations through a better understanding of biologically relevant inter- and intra-molecular interactions. Also, more efficient sampling methods are needed that can quickly and accurately identify conformational states of structures that are related to functions of biological interest.

The function of a protein depends on its structure which can be affected by interactions with other molecules, especially small chemicals. Modeling such interactions between large and small molecules is an important area of research in molecular discovery that is being used to treat diseases and prevent toxicity. Research in this area is complicated by structural modifications that occur during the metabolism of chemicals. Changes in chemical structure via metabolism affect subsequent

interactions within biological pathways of multiple proteins. Improvements in the accuracy and speed of predictive and experimental techniques that capture both chemical metabolism and activity of small chemicals though multi-protein pathways are needed to accommodate the study of many molecules within biological organisms and their environments.

Scope of Dissertation

Molecular structures and their interactions are basic components of life that continue to guide researchers to new methods for disease treatment and prevention. In theory, every disease or adverse health effect can be understood at the lowest atomic-level of interactions that occur between and within molecules. The research presented in this dissertation deals with ongoing challenges needed to improve the state of structure-based molecular modeling with respect to interactions within and between protein and chemical structures. In particular, this research extends the usefulness of structure-based molecular modeling methods to sample the unresolved structure of a protein intermediate state and its chemical interactions related to a disease, predict chemical interactions and metabolism between multiple proteins in a toxicity pathway, and describe a new noncovalent interaction that can be used to better approximate the thermodynamic properties of molecular structures.

Introduction to Virtual Docking

The study of interactions that control molecular structure and function has lead to understanding and treating diseases like HIV[3], breast cancer[4], and obesity[5]. Despite much success in the field of structure-based discovery, current molecular modeling techniques need to still improve in order for the field to reach its anticipated potential of describing in detail how any given biological function arises from molecular structures and their interactions. It is unlikely that all structures and interactions will ever be completely modeled using experimental methods due to the vast number of structures, physical limitations of studying some important structures, and complexity of understanding increasingly higher-orders of structure. Predictive computational methods that simulate molecular interactions can aid experiments in the number of structural interactions that can be studied. The continued application, development, and integration with experiments of structure-based simulation techniques is an important area of research in molecular discovery and is the main focus of this dissertation.

Virtual docking is a computational modeling technique that allows for the structure and thermodynamic properties (i.e., binding) between two or more molecules of known structure to be predicted by their molecular interactions. A typical docking simulation can be run in a matter of seconds on a single processor, and the scaling potential of docking allows for straightforward parallelization across high-performance computing clusters and even supercomputing platforms. The usefulness of this technique is in the low cost and time efficiency of performing computations compared to carrying out physical experiments. Docking can be thought of as a virtual microscope to use in complement with experiments or when physical methods are simply unavailable or impractical.

Kuntz first described the basic implementation of rigid docking in 1980[6], demonstrating that molecular geometries between independent rigid molecules could be sampled algorithmically to reproduce experimentally known binding poses. Since then docking has been increasingly used as a tool to study binding between small drug-like molecules, proteins, lipids, nucleic acids, and carbohydrates[7-11]. Rigid docking is based on the "lock and key" theory of binding proposed by Fischer in the late 1890s[12] and is the most basic and widely used implementation of docking. Flexible docking is based on the "induced fit" theory proposed by Koshland in the early 1960s[13]. Docking results are often improved through flexible docking schemes, but such methods require better theoretical understanding about molecular interactions and require more computational time. Relatively fast flexible ligand methods have been used since the mid-1980s[14,15]. Localized protein sidechain flexibility can be simulated much like the induced fit scheme[16,17]. In some cases, protein backbone flexibility can be simulated[18,19], but these most advanced techniques are limited in the sampling of phase space for large macromolecules such as proteins.

The quality of a docking model is evaluated by its ability to reproduce experimental binding geometries, predict experimental binding affinities, and distinguish between molecules that bind and do not bind to one another (i.e., binders and non-binders). The basic docking algorithm involves sampling spatial arrangements between two molecules (i.e., generating "poses") and then determining the fitness (i.e., "scoring") for each bound configuration. Pose generation can involve varying degrees of flexibility for the atoms of a molecular system and also hierarchical levels of scoring. Knowledge-based or empirically derived scoring methods, e.g., basic shape and chemical matching, are often used to quickly reduce the quantity of poses that are to be further evaluated with more flexible and complex scoring methods such as those involving force fields. Force fields allow for approximating free energy changes for varying spatial arrangements of atoms in a molecular system. Many force field energy functions exist to describe both bonded and non-bonded terms that are thought to be important for molecular stability, e.g., van der Waals, electrostatics, desolvation, bonds, and angles. Inclusion of other relevant energy terms in force fields is an active area of research for improving molecular modeling.

Why Use Predictive Docking?

Docking has been used as a popular tool for structure-based virtual screening[20,21]. Virtually screening of chemical libraries aims to find new chemical architectures that can modulate the structure and function of proteins involved in disease and/or toxicity. The scaling and efficiency of docking allows for virtual querying of large datasets. Libraries available for virtual screening can be in the order of millions of compounds which is something that experiments cannot test on their own. Virtual screening is likewise aided by experiments which provide an initial set of small molecules with known activities to be used for optimizing a model's predictive accuracy. This process involves choice of initial structures, degree of structure flexibility, robustness of sampling procedures for generating new poses, and choice of scoring methods to evaluate the energetic likelihood of new poses. Once optimized to agree with experiments, a docking model can then be used to more reliably predict binding

affinities for new compounds with unknown activities. This computational prioritization and experimental validation scheme lowers the costs and time associated with testing a large number of compounds. Docking has traditionally been used to find chemicals that bind to a single protein target, but the concept of reverse screening is also possible whereby many proteins are assessed for their binding to a single or several chemical targets[22].

Over 100,000 three-dimensional structures have been deposited in the Protein Data Bank (PDB)[23] and the growth of this database is correlated with the growth of molecular docking since structure is essential for docking. In the post-genomic era, the PDB continues to grow by thousands of entries each year. Alongside this growth is an increasing pool of sequenced genes and proteins that can be paired with homologous protein structures from the PDB to build homology models which are suitable for docking[24,25]. Homology models are validated for use in virtual screening in the same way as experimental structures, i.e., by showing their predictive power on a training set of chemicals with known binding affinities.

Despite the current usefulness and success of docking, there is still a need for improving its accuracy, scaling, and application. The work within this dissertation provides novel applications and developments which improve the field of molecular docking and structural modeling in general. It is demonstrated that docking can be integrated with *in vitro* and other *in silico* techniques to overcome several structural modeling challenges such as identifying an intermediate state structure that is experimentally unresolved, and modeling chemical interaction and metabolism in multi-protein pathways. In addition, a better understanding of an unconventional noncovalent interaction between aromatic and anionic functional groups in biological molecules is presented for improving protein and ligand modeling.

Introduction to Chapter 1: Modeling of an Intermediate State Structure

Chapter 1 reports an application that uses virtual docking to create a model that predicts the binding of chemicals to a single protein target of medicinal interest. The specific target, alpha-1-antitrypsin, represents a traditional single-protein application of docking; however it also presents a unique docking challenge since the desired conformation of the protein structure is a theoretical intermediate state for which no experimental structure exists. The targeted intermediate state is a suspected transition structure between two stable states which are represented in the PDB, so a unique homology modeling strategy is used to merge the two stable state structures together in such a way as to represent the desired intermediate state. The resulting homology-based docking model is validated using an initial training set of compounds with known experimental binding affinities, and the modeled structure is then applied to virtually screen a chemical library for new drug candidates to treat the disease associated with alpha-1-antitrypsin deficiency.

Introduction to Chapter 2: Multi-Protein Modeling and Metabolism Prediction

An important aspect of understanding biological function is modeling the interactions and structural changes that occur to a chemical during cellular transport through multi-protein pathways which ultimately affect a compounds bioavailability, bioactivity, and toxicity. Chapter 2 of this dissertation presents a biologically significant application of docking which involves integrating several docking models with other predictive and experimental techniques in order to model the metabolism and augmented bioactivity of an environmental toxin, PCB-30. The resulting model correctly predicts the metabolism of PCB-30 by two cytochrome P450 enzymes and the relative bioactivities for the primary metabolites with the estrogen hormone receptor.

Introduction to Chapter 3: Anion-pi Interactions in Molecular Modeling

Biological structures and functions are the result of a complex network of weak and strong molecular interactions within and between the constituent atoms of molecules. Virtual docking works on the principle of being able to account for the relevant interactions which affect the three-dimensional structure and binding between molecules. In chapter 3 of this dissertation, the theory of anion-pi interactions, an emerging noncovalent interaction, is introduced along with work demonstrating the significance of anion-pi interactions in biological protein and protein-ligand structures from the PDB. This work is organized into three sections (3.1, 3.2, 3.3). The results of this work may lead to the development of additional forcefield energy terms for improving protein and ligand modeling.

References

(1) Borhani, D.W., and Shaw, D.E. 2012. The future of molecular dynamic simulation in drug discovery. *J Comput Aided Mol Des*. 26, 15–26.

(2) Lodish, H., Berk, A. Zipursky, S.L., et al. 2000. *Molecular cell biology. 4th edition.* New York: W.H. Freeman. Section 1.2, The molecules of life.

(3) Lee-Huang, S., Huang, P.L., Zhang, D., Lee, J.W., Bao, J., Sun, Y., Chang, Y-T., Zhang, J., Huang, P.L. 2007. Discovery of small-molecule HIV-1 fusion and integrase inhibitors oleuropein and hydroxytyrosol: l. fusion inhibition. *Biochem Biophys Res Commun.* 354, 872-878.

(4) Jiang, Q., Zheng, S., Wang, G. 2013. Development of new estrogen receptor-targeting therapeutic agents for tamoxifen-resistant breast cancer. *Future Med Chem.* 5, 1023-1035.

(5) Shyh, G., Cheng-Lai, A. 2014. New antiobesity agents: lorcaserin (Belviq) and phentermine/topiramate ER (Qsymia). *Cardiol Rev*. 22, 43–50.

(6) Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., Ferrin, T.E. 1982. A geometric approach to macromolecule-ligand interactions. *J Mol Biol.* 161, 269–288.

(7) Sousa, S.F., Fernandes, P.A., Ramos, M.J. 2006. Protein-ligand docking: current status and future challenges. *Proteins.* 65, 15-26.

(8) Gray, J.J. 2006. High-resolutation protein-protein docking. *Curr Opin Struc Biol.* 16, 183-93.

(9) Holt, P.A., Chaires, J.B., Trent, J.O. 2008. Molecular docking intercalators and grove-binders to nucleic acids using Autodock and Surfex. *. Chem Inf Model.* 48, 1602-15.

(10) Wang, T., Zhou, X., Wang, D., Yin, D., Lin, Z. 2012. Using molecular docking between organic chemicals and lipid membrane to revise the well known octanol-water partition coefficient of the mixture. *Envron Toxicol Pharmacol.* 34, 59-66.

(11) Sapay, N., Nurisso, A., Imberty, A. 2013. Simulation of carbohydrates, from molecular docking to dynamics in water. *Methods Mol Biol*. 924, 469-483.

(12) Fischer, E. 1890. Ueber die optischen Isomeren des Traubenzuckers, der Gluconsäure und der Zuckersäure. *f3er. L)tsch. Chcm. Grs.* 23, 2611.

(13) Thoma, J.A., Koshland, D.E. 1960. Competitive inhibition by substrate during enzyme action. Evidence for the induced-fit theory. *J Am Chem Sot.* 82, 3329-33.

(14) DesJarlais, R.L., Sheridan, R.P. , Dixon, J.S., Kuntz, I.D., Venkataraghavan, R. 1986. Docking flexible ligands to macromolecular receptors by molecular shape. *Journal of medicinal Chem.* 29, 2149-53.

(15) Lorber, D.M., Shoichet, B.K. 1998. Flexible ligand docking using conformational ensembles. *Protein Sci.* 7, 938-50. *J Mol Bio.* 235, 345-356.

(16) Leach, A.R., 1994. Ligand docking to protein with discrete side-chain flexibility.

(17) Schueler-Furman, O., Wang, C., Baker, D. 2005. Progress in protein-protein docking: atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility. *Proteins.* 60, 187-94.

(18) Knegtel, R.M.A, Kuntz, I.D., Oshiro, C.M. 1997. Molecular docking to ensembles of protein structures. *J Mol Biol.* 266, 424-440.

(19) Claussen, H., Buning, C., Rarey, M., Lengauer, T. 2001. FlexE: efficient molecular docking considering protein structure variations. *J Mol Biol.* 308, 377-95.

(20) Shoichet, B.K., McGovern, S.L., Binqing, W., Irwin, J.J. 2002. Lead discovery using molecular docking. *Curr Opinion in Chem Biol.* 6, 439-446.

(21) Schneider, G., Bohm, H.-J. 2002. Virtual screening and fast automated docking methods. *Drug Discovery Today.* 7, 64-70.

(22) Kharkar, P.S., Warrier, S., Gaud, R.S. 2014. Reverse docking: a powerful tool for drug repositioning and drug rescue. *Future Med Chem.* 6, 333-342.

(23) Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., and Zardecki, C. 2002. The protein data bank, *Acta Crystallogr. D Biol. Crystallogr. 58*, 899-907.

(24) Kairys, V., Fernandes, M.X., Gilson, M.K. 2006. Screening drug-like compounds by docking to homology models: a systematic study. *J Chem Info Model.* 46, 365-79.

(25) Rockey, W.M., Elcock, A.H. 2006. Structure selection for protein kinase docking and virtual screening: homology models or crystal structures? *Curr Protein Pept Sci.* 7, 437-57.

CHAPTER 1.1.
DISCOVERY OF A POTENT INHIBITOR OF Z-ALPHA1 ANTITRYPSIN POLYMERIZATION

A version of this chapter was submitted for publication by Valerie Berthelier and Jason B. Harris, Kasey Estenson, Jerome Baudry:

Valerie Berthelier, Jason B. Harris, Kasey Estenson, Jerome Baudry. "Identification of S-(4-nitrobenzyl)-6-thioguanosine as Inhibitor of Z-Alpha1 Antitrypsin Polymerization". *Biochemistry* (2014). (recently submitted)

Figure 1.1.9 represents additional work not included in the recently submitted paper. The work and writing in this article was equally contributed to by Jason B. Harris and Valerie Berthelier. V. Berthelier designed and carried out experimental protocols. J.B. Harris designed and carried out all computational models. J.B. Harris and V. Berthelier co-drafted the manuscript. K. Estenson assisted in carrying out experiments. J. Baudry and V. Berthelier served as faculty mentors and assisted in manuscript revisions.

Abstract

Polymerization of the Z variant alpha-1-antitrypsin (Z-α1AT) results in the most common and severe form of form of α1AT deficiency (α1ATD), a debilitating genetic disorder whose clinical manifestations range from asymptomatic to fatal liver or lung disease. As the altered conformation of Z-α1AT and its attendant aggregation are responsible for pathogenesis, the polymerization process *per se* has become a major target for the development of therapeutics. Based on the ability of Z-α1AT to aggregate by recruiting on its s4A cavity the reactive center loop (RCL) of another Z-α1AT, we developed a high-throughput screening assay that uses a modified 6-mer peptide mimicking the RCL to screen for inhibitors of Z-α1AT polymer growth. A subset of a commercially available library of small compounds with MWs ranging from 300 to 700 Da was used to test the assay's capabilities, and the inhibitor S-(4-nitrobenzyl)-6-thioguanosine was found. To validate S-(4-nitrobenzyl)-6-thioguanosine, an *in silico* strategy was pursued and the intermediate α1AT M* state modeled to allow molecular docking simulations, which explore various potential binding sites. Docking results predict that S-(4-nitrobenzyl)-6-thioguanosine can bind at the s4A cavity or at the edge of β-sheet A. The former binding site would block RCL insertion whereas the latter site would prevent β-sheet A from expanding between s3A/s5A, and thus indirectly impede RCL binding. Altogether, our investigations have revealed a novel compound that specifically inhibits the formation of Z-α1AT polymers, as well as *in vitro* and *in silico* strategies for identifying small molecules for treatment of α1ATD.

Introduction

Human α1-antitrypsin (α1AT) is the most abundant member of the serine protease inhibitor (SERPIN) family. It is a soluble 52-KDa glycoprotein synthesized primarily by hepatocytes and delivered to the lungs to accomplish its critical function: inactivation of the proteinase neutrophil elastase (NE), a mediator of alveolar destruction.[1] Defective folding, trafficking and secretion into the plasma of α1AT are responsible for α1AT deficiency (α1ATD).[2,3]

The structural flexibility of α1AT is important for it to perform its anti-protease function and ensure lung integrity. With a core domain composed of 3 β-sheets A, B and C, and 9 α-helices, α1AT features an exposed and flexible reactive center loop (RCL) that serves as bait for NE. Upon binding to the proteinase, a dramatic conformational change occurs as RCL is cleaved and translocates into β-sheet A to form the new central and fourth strand, s4A. The translocation event carries along NE from one side to the other of α1AT, causing its inactivation by forming an irreversible, higher molecular weight suicide complex.[4,5] A reduction or lack of this inhibition through loop-sheet insertion and proteolytic cleavage is thought to be the underlying mechanism responsible for α1ATD.[6,7]

Over 100 genetic variants of α1AT have been identified with the Z-type being responsible for the most common and severe form of the disease in homozygous patients[8]. The punctual mutation E342K in Z-α1AT renders the anti-protease prone to aggregation and unable to be secreted into the blood stream resulting in a 90% decrease in NE inhibition within the lungs. Accumulation of polymers of Z-α1AT in the endoplasmic reticulum (ER) of hepatocytes leads to proteotoxic stress and associated liver diseases.[9–11] In addition to sequestration of polymers in the ER of hepatocytes, the E342K mutation has two additional disease-causing effects. It causes Z-α1AT to be 5-fold less effective in accomplishing its inhibitory function[12,13] and it promotes the spontaneous formation of Z-α1AT polymers within the lungs, thereby further reducing the already depleted levels of α1AT that are available for alveola protection.[14] Moreover, the conversion of Z-α1AT from a monomer to a polymer renders it a chemoattractant for human neutrophils.[15,16] To summarize, emphysema associated with Z-α1ATD results from a combination of (1) loss of function of the anti-protease, which leads to the absence of circulating α1AT, decrease of its inhibitory activity, and intra-alveolar polymerization, and (2) gain of toxic function from the neutrophil chemotactic properties of intra-alveolar polymers.

Preventing formation and accumulation of Z-α1AT polymers could be crucial to treat α1ATD.[17] For this reason, the mechanisms by which Z-α1AT form polymers have been under intense investigation. As the substitution of the glutamic acid residue at position 342 by a lysine provokes a perturbation in the native structure by opening the β-sheet A, biochemical evidence reveals the formation of an unstable and polymerogenic intermediate M* with its own RCL partially inserted.[18] The opening of the s4A cavity allows the creation of a sequential β-strand linkage between the RCL of one serpin and β-sheet A of another, leading to the formation of a dimer and then polymers.[6,19–21] Additional models for Z-α1AT polymerization have also recently been proposed based on the crystal structures of a dimer of the serpin antithrombin[22] and a trimer of a disulfide mutant of α1AT,[23] suggesting that assembly pathways of Z-α1AT could be diverse and therefore arising from structurally and/or dynamically distinct polymerogenic intermediates.

Various strategies have been pursued in order to prevent or even attenuate Z-α1AT polymerization such as increasing the mutant protein secretion with the use of

osmolytes[24–26] or by blocking Z-α1AT polymerization by either filling the s4A cavity with peptides[18] or crowding another hydrophobic pocket of Z-α1AT with small compounds screened virtually.[27] While extensive progress has been made, none of these strategies has been entirely successful so far. To achieve this goal, we developed a set of novel and integrated *in vitro* and *in silico* screenings methods; the *in vitro*, being a high-throughput screening assay using a modified small peptide previously reported as a s4A cavity filler,[18] and the *in silico* being a virtual docking model able to predict and help rationalize the binding of compounds to α1AT, including in the S4A cavity. Here, we present how using these two combined methods we were able to identify, rationalize and validate S-(4-nitrobenzyl)-6-thioguanosine as a specific inhibitor of Z-α1AT polymerization.

Materials and Methods

General Materials and Methods

The peptide acetyl-FLEAIGGG-Q-GKKG containing the 6-mer sequence of the RCL was synthesized by custom solid-phase from the Keck Biotechnology Center at Yale University (http://info.med.yale.edu/wmkeck/). A biotinylated version of the peptide (bPEG-peptide) was obtained by appending a biotinyl-polyethylene glycol spacer on the □-amide group of the glutaminyl residue. The presence of the Lys residues confer a positive net charge to the peptide at neutral pH, enhancing its general solubility. The wild type and Z-α1AT proteins, prepared according to published protocol,[28] were graciously provided at a concentration of 1 mg/ml by Professor Lomas, Cambridge Institute for Medical Research, University of Cambridge, UK, and stored at 4 ºC.

The rabbit anti-human α1AT antibody (serum fractions IgG) was purchased from Abcam, Cambridge, MA.

The test group RK-001 of the LOPAC library (Library of Pharmacologically Active Compounds, Sigma-RBI, Natick, MA) containing 80 lyophilized chemical compounds was prepared in a 96-well plate format. All compounds were resuspended in 2 ml DMSO at a concentration of approximately 4 mM, based on an estimated MW average of 500 g/moles, and stored at 4 ºC.

Preparation of the bPEG-peptide

The synthesized bPEG-peptide was first solubilized in 50% formic acid at a concentration of ~1mg/ml, injected onto a Zorbax C3 Column and purified by RP-HPLC at a rate of 4ml/min. The resulting purified peptide was then lyophilized, resuspended into H_2O and stored at -20 ºC. After amino acid analysis of the peptide (Commonwealth Biotechnologies Inc., Richmond, VA), various amounts were injected onto RP-HPLC in order to establish a standard curve, allowing us to determine the exact concentration of each new batch of purified peptide that we prepared.

Preparation of Working Compound Plates

LOPAC compounds were transferred from their original 96-well plates to new 96-well working plates with low evaporation lid (BD Falcon plates non treated, Becton

Dickinson Labware, San Jose, CA) in respect to their initial location, and adjusted to a concentration of 1 mM in PBS 1X containing 50% DMSO. First and last columns were filled only with PBS 1X/DMSO (50/50). Working plates were sealed with an adhesive overlay, covered and stored at 4º C until further utilization.

Set up of the Microplate Screening Assay

The assay is based on the principle of a competitive ELISA.[29] Wells were coated by passive adsorption with a 1/1000 solution in PBS 1X of capture α1AT Ab. The screening microplate was sealed with an adhesive overlay and incubated for 2 h at 37 ºC. The wells were then washed three times with extension buffer (PBS 1X and 0.01% Tween 20), blocked for 1 h at 37 ºC with 0.3 % gelatin and washed again. Screening results described in this paper were carried out with screening microplates freshly made. However, screening microplates can be filled with PBS 1X, hermetically sealed and stored at 4 ºC for one week prior to use.

Z-α1AT Polymerization Inhibition Assay

In parallel with the preparation of the microplate screening assay, polymerization reactions were carried out in 96-well plates with or without the LOPAC small compounds. A 100 molar excess of bPEG-peptide was used with 4 μg/well of Z-α1AT.

Each polymerization reaction plate was organized as follows: the first column contained only bPEG-peptide (background control); the second to eleven columns contained Z-α1AT, compounds and bPEG-peptide; and the last column contained Z-α1AT, bPEG-peptide and no compound (reaction control). Assay wells were set up by adding to each well 20 μl of protein and 20 μl of compound from a working plate. After 3 min, 160 μl of bPEG-peptide at 48 μM was added. The plate was then sealed, shaken on a microplate shaker gently for 5 s to ensure homogeneity of the different reactants, and placed at 37 ºC for 16 h. All wells contained 5% DMSO.

At the end of the 16 h incubation time, 100 μl from each well were transferred into the corresponding well of the microplate screening assay. One hour later, the screening plate was washed three times and then incubated in the dark for 1 h at room temperature with 100 μl/well of 1 ng/μl of europium streptavidin (Perkin Elmer, Boston, MA) in 0.5% BSA-extension buffer. Three final washes in extension buffer were carried out and the europium was released from streptavidin by the addition of 100 μl of enhancement solution (Perkin Elmer). After 5 min, europium fluorescence was measured by time-resolved fluorometry in a Victor 2 counter (Perkin Elmer) and then converted to fmoles of bPEG-peptides recruited into Z-α1AT. Assays were conducted in triplicate by processing three identical plates in parallel.

Determining IC_{50} Values of Inhibitors

As described above, 4 μg/well of Z-α1AT were incubated for 3 min with various concentrations of a compound identified as an inhibitor, the highest concentration starting at 400 μM. The concentrations of the compounds were revised according to the

true MW of the molecule. Following the 3 min incubation, 160 μl of 48 μM bPEG-peptide were added and the rest of the protocol was applied as described above.

Z-α1AT Polymerization Sedimentation Assay

A solution of 0.1 mg/ml of Z-α1AT in PBS 1X was incubated at 37 ºC with or without 100 μM of S-(4-nitrobenzyl)-6-thioguanosine. Progress of the polymerization reaction was followed by quantitative RP-HPLC on centrifugation supernatants (20 min, 20,000 × *g*) of reaction aliquots. Quantitative determination of the Z-α1AT monomer disappearance was calculated according to a pre-established standard curve.

Preparing Protein Structures for Homology Modeling and Docking Simulations

The crystal structures of the mutant Z-α1AT (PDB code: 3T1P) and the wild type M-α1AT (PDB codes: 3CWM and 1QLP) were obtained from the RCSB Protein Database.[23,30,31] Initial preparation of the receptor structures was carried out with the program MOE[32] (Molecular Operating Environment). Co-crystallized water molecules were deleted from both structures. For the polymerized mutant (3T1P), only the first monomer was retained and the s4A binding cavity was created between the s3A/s5A by deleting residues 345-356, which correspond to the inserted residues of the RCL. The protonation state of atoms was assigned using Protonate 3D[33] utility in MOE at pH 7, 300 K and 0.1 M salt concentration. Solvent effects were implicitly included by using a distance-dependent dielectric. Partial charges were assigned to receptor atoms using MMFF94s[34] force field parameters as implemented in MOE.

Homology Modeling Procedure

Modeling of the M* intermediate state as described below was performed using the Homology Model facility in MOE. The wild type crystal structure of α1AT (PDB code 1QLP) was used as a template for modeling i) the position of the RCL when not inserted into β-sheet A and ii) the position of the c-terminal loop within β-sheet B when it is not participating in a domain swap. The polymerized Z-mutant structure (PDB code: 3T1P) was used to model the expanded position of β-sheet A but omitting s4A to leave a cavity between s3A/s5A where the RCL would otherwise be found. Fragments from each template structure were joined at transition points selected by superimposing the structures and choosing those residues between fragments with near overlapping atom positions. A total of 25 homology models were generated with unique carbon backbone positions, and for each of those 25 models, 5 additional models (*i.e.* a total of 125 models) were created with alternate side chain positions. These initial models were energy minimized to a gradient of 0.1 kcal/mol·Å. A final M* model (Model 126) was created using the Generalized Born / Volume Integral (GB/VI) energy scoring method[35] to select the best initially packed structure and then further energy minimizing it to a gradient of 0.01 kcal/mol·Å.

Docking Simulation Procedure

Three-dimensional structures of the 80 *in vitro* tested chemicals**,** including S-(4-nitrobenzyl)-6-thioguanosine, were obtained in SDF format from the electronic LOPAC library, test group RK-001. Partial charges were added to each ligand atom using MMFF94s forcefield parameters and the structures were energy minimized to a gradient

of 0.1 kcal/mol·Å. The energy-minimized ligands were docked into three structural variations of α1AT: M* intermediate (Homology Model 126), mutant Z-α1AT (PDB: 3T1P) and wild type M-α1AT (PDB: 3CWM) using the docking function built into MOE. Binding sites were identified using the MOE Site Finder facility. Separate docking simulations of the 80 compounds were carried out for each receptor and at each potential binding site. Initial placement of ligand atoms was done with the Triangle Placement method (seeds in 3 atoms at time). The predicted free energy of binding for each initially docked ligand pose was calculated using the London dG^{32} scoring method from within MOE. The top five scoring poses were further energy minimized using the MMFF94s force field, allowing ligand atoms and protein side chains within 6 Å of each docked ligand to be treated as flexible. A tethering weight of 10 kcal/mol/$Å^2$ was applied to partially restrain flexible atoms around their original location. A final docking score for each energy-minimized pose was calculated using the Affinity dG^{32} scoring method.

Virtual Screening the NCI Diversity Database

The M* Model was used to assess the binding of 1596 compounds from the September 2013 NCI Diversity Set (http://dtp.nci.nih.gov/branches/dscb/div2_explanation.html) and the original 80 compounds from the LOPAC library, test group RK-001. The same methods for docking preparation and scoring found in the Docking Simulation Procedure were used in the virtual screening.

Results

Principal Characteristics of the Z-α1AT Polymerization Inhibitor Screening Assay

Previous studies have shown that a 6-mer peptide whose amino acid sequence contains the RCL sequence FLEAIG can specifically bind the Z-mutant at its opened s4A pocket, but not the wild type. [18] The Z-α1AT high-throughput microplate screening assay is based on this concept. Thus, we designed a similar peptide and added a biotin-polyethylene glycol (bPEG) tag at the C_{term} of the reactive loop sequence as well as some hydrophilic amino acids to increase peptide solubility. The insertion of a PEG-based spacer prevents possible steric hindrance between the peptide and the biotin molecule, resulting in better avidin binding and therefore, a more accurate measurement of the biological activity.

To assess the ability of small molecules to inhibit the recruitment of the bPEG-peptide into Z-α1AT, microtiter plate wells containing attached Z-α1AT are subjected for 3 min to library compound before addition of bPEG-peptide. The amount of bPEG-peptides incorporated into the mutant protein is then determined by a europium-streptavidin treatment and time-resolved fluorescence measurements. The inhibition effect of a compound is calculated as a percentage with respect to a reaction control – i.e. Z-α1AT that has only been exposed to the biotinylated peptide and not to a compound. Any compound showing an inhibitory effect of at least 50% is considered as a hit.

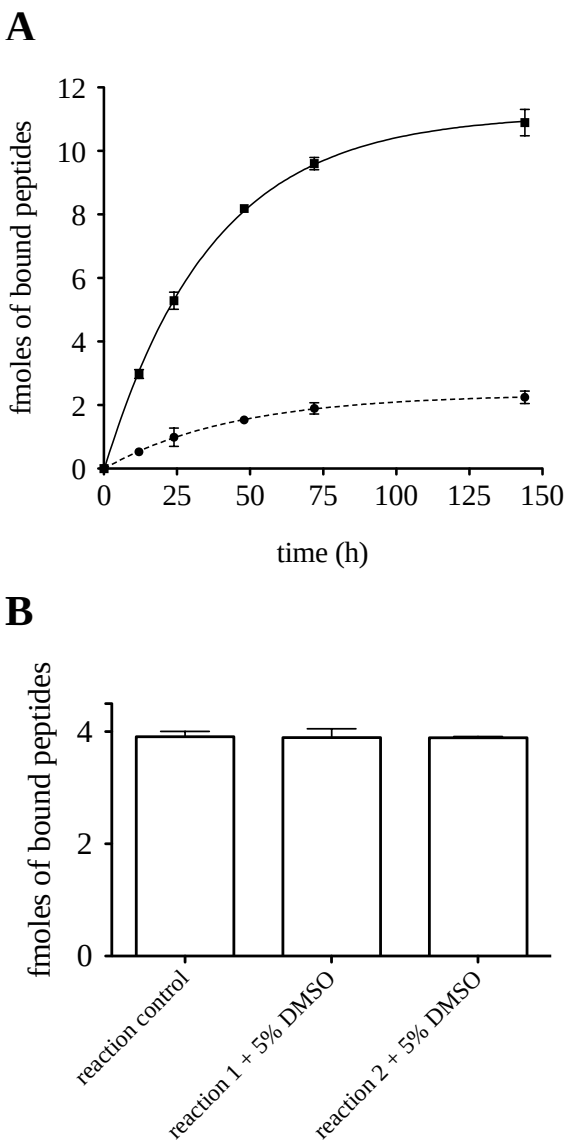


Figure 1.1.1 Kinetic diagram of bPEG-peptide binding to α1AT. (A) Four micrograms per well of attached (■) Z-α1AT or (●) M-α1AT were incubated for various times in presence of 38.4 μM bPEG-peptide. (B) Z-α1AT was incubated in presence of 5% DMSO and bPEG-peptide for 16 h. Errors bars reflect the standard deviation of three replicates.

Regarding its ability to bind Z-α1AT, we found the bPEG-peptide association kinetics to be in favor of the mutant proteinase with an initial association rate of 0.22 ± 0.08 fmoles·h^{-1} *vs.* 0.042 ± 0.1 fmoles·h^{-1} for the wild type (Figure 1.1.1A). We also found that an incubation period of 16 hrs for the peptide with Z-α1AT is an adequate screening end-point for the screening assay as this time period is associated with a high signal-to-noise ratio. In addition, the presence of 5% DMSO in the wells does not affect

the bPEG-peptide binding kinetics (Figure 1.1.1B). Since compound libraries are generally stored in DMSO, this feature makes the assay well suited for a high-throughput screening assay

Finally, this screening assay exhibits very good reproducibility as reflected by the error bars shown in Figure 1.1.2. It requires only small amount of protein and low concentrations of bPEG-peptide, which make it both economical and physiological.

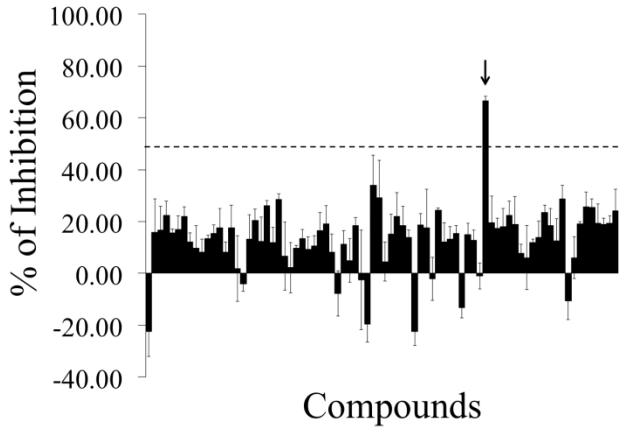


Figure 1.1.2. Pattern of inhibition resulting from the screening of 80 unknown LOPAC compounds. A 96-well plate was coated with 4 μg/well of Z-α1AT and incubated for 16 h with 100 μM of various compounds and 38.4 μM of bPEG-peptide. The black arrow indicates the compound that corresponds to S-(4-nitrobenzyl)-6-thioguanosine and gives an inhibition effect of 67 ± 2 % and. The error bars are the standard deviation of three individual experiments.

S-(4-nitrobenzyl)-6-thioguanosine Identified as Inhibitor of Z-α1AT Polymerization

The test group RK-001 (80 compounds) of the small commercially available LOPAC library containing drug-like molecules was used to test the performance of the screening assay. Figure 1.1.2 shows a typical screening result. As indicated in the figure, only one compound of the tested compound plate appears as a hit, exhibiting a 67 ± 2 % inhibition activity at 100 μM. This compound is S-(4-nitrobenzyl)-6-thioguanosine. To confirm its ability to inactivate Z-α1AT polymerization, dose-responses curves were carried out and an IC_{50} of 73 ± 0.12 μM calculated (Figure 1.1.3A). The IC_{50} value obtained is in the micromolar range and matches well with the screening results.

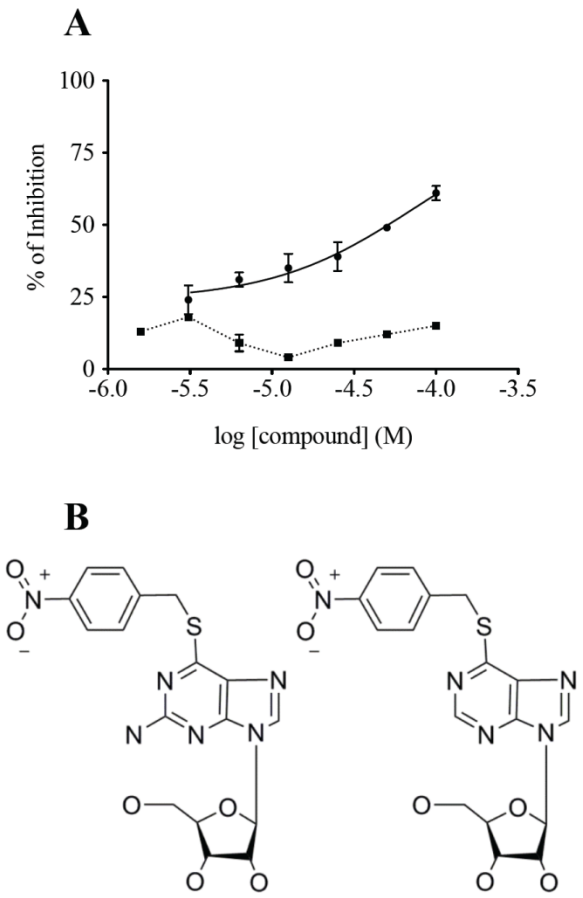


Figure 1.1.3. S-(4-nitrobenzyl)-6-thioguanosine inhibits bPEG-peptide binding to Z-α1AT. (A) Dose-response curves were assayed for various concentrations of (●) S-(4-nitrobenzyl)-6-thioguanosine and (■) its homologue S-(4-nitrobenzyl)-6-thioinosine. (B) Chemical structures of (left) S-(4-nitrobenzyl)-6-thioguanosine and (right) S-(4-nitrobenzyl)-6-thioinosine. The errors bars are the standard deviation of an experiment conducted in triplicate.

To define a better pharmacophore, and therefore to identify any additional structural element required for inhibiting Z-α1AT polymerization, we then compared our compound to the entire database that regroups all of the LOPAC molecules. Interestingly, we found that S-(4-nitrobenzyl)-6-thioanosine possesses a very similar structure, differing by a single amino group, but did not show any inhibitory effect, neither during the original screening nor in the validation assay (Figures 1.1.3A and 1.1.3B).

Validation of the Action of S-(4-nitrobenzyl)-6-thioguanosine

A polymerization reaction was set up in presence or absence of 100 µM of S-(4-nitrobenzyl)-6-thioguanosine and the disappearance of the Z-α1AT monomer monitored

by RP-HPLC – a diminution in monomer concentration indicates that the protein has been recruited into polymers. We found that Z-α1AT has its polymerization rate decreased 33 times in presence of the compound and that its effect is long lasting (Figure 1.1.4).

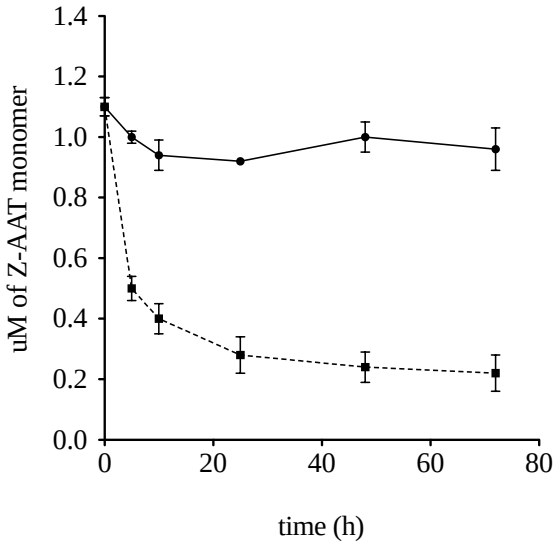


Figure 1.1.4. Effect of S-(4-nitrobenzyl)-6-thioguanosine on Z-α1AT polymerization. The protein was incubated with (●) or without (■) 100 μM of S-(4-nitrobenzyl)-6-thioguanosine for various time at 37 ºC. The error bars are the standard deviation of three separate experiments.

An additional approach to validate the action of S-(4-nitrobenzyl)-6-thioguanosine was to carry out an *in silico* molecular modeling and simulation strategy with the intent to gain mechanistic insights into how this small molecule may interact with the protein structure and prevent polymerization. Therefore, virtual docking and homology modeling were used to explore hypothetical binding sites and their putative molecular interactions.

Structural Modeling of α1AT and the M* Intermediate

In order to investigate all of the potential binding sites of S-(4-nitrobenzyl)-6-thioguanosine on α1AT, including the ones located in the s4A cavity at the RCL insertion site, we used two PDB crystal structures, 1QLP and 3T1P, which respectively correspond to the wild type M-α1AT and polymerized Z-α1AT states. However, as the s4A cavity does not exist in any crystal structure of α1AT, a theoretical model comparable to M* had to be created. The M* intermediate state is described to have the following three structural features: i) an expanded β-sheet A with a s4A cavity between s3A/s5A, ii) an RCL at the precipice of inserting between s3A/s5A, and iii) the C_{term} loop

inserted within β-sheet B and not participating in a domain swap with another protein. These important features of the M* model are represented in the homology model built from the two available crystal structures of α1AT (Figure 1.1.5).

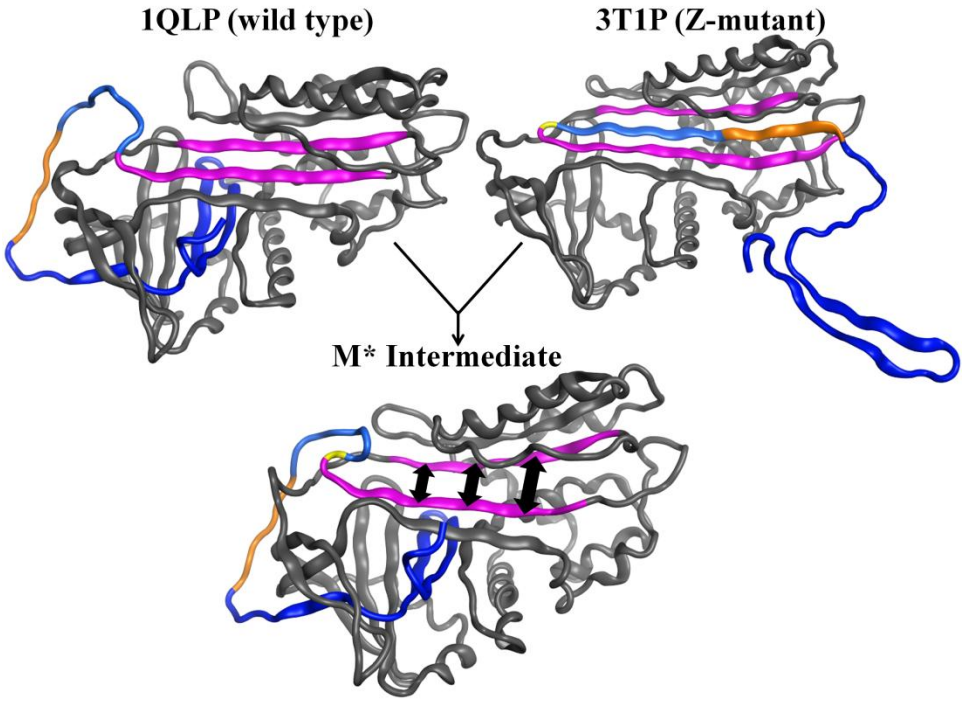


Figure 1.1.5. The three models of α1AT protein. (Top left) Structure of wild type (PDB: 1QLP) with the RCL not inserted and β-sheet A not expanded. (Top right) Structure of Z-mutant (PDB: 3T1P) with the RCL inserted and β-sheet A expanded. (Bottom middle) Intermediate M* model with an expanded β-sheet A (retained from structure 3T1P), RCL not inserted into the RCL cavity (retained from structure 1QLP), and C_{term} loop inserted into β-sheet B (retained from structure 1QLP). (Purple) Strands 3 and 5 from β-sheet A. (Dark blue) C_{term} loop within β-sheet B. (Light blue) RCL. (Orange) Residues of the RCL corresponding to the analogous 6-mer peptide. (Black arrows) s4A cavity.

To build the M* state homology model, a total of five protein fragments of the two crystal structures, 1QLP and 3T1P, were merged. Figure 1.1.6 shows that fragment 1 consists of residues 1-105 (1QLP) which model the right side of β-sheet B, with respect to beta strands adjacent to the right side of the C_{term} loop. Fragment 3 consists of residues 205-291 which constitute the left side of β-sheet B, with respect to beta strands adjacent to the left side of the C_{term} loop. Together these two fragments model the position of β-sheet B so that the RCL residues from fragment 5 (residues 345-394 from

1QLP) can be placed on the outside of the s4A pocket along with the C_{term} loop buried within β-sheet B. The position of strands s1A, s2A, and s3A in β-sheet A are modeled from fragment 2 (residues 106-204 from 3T1P), and fragment 4 (residues 292-344 from 3T1P) models the position of strands s5A and s6A. Together, the positions of fragments 2, 4 and 5 create the cavity s4A between s3A/s5A, which would otherwise be the site of RCL insertion. This opened conformation of α1AT represents one of the possible structures of the unstable M* intermediate state for which experimental methods such as crystallography cannot reproduce.

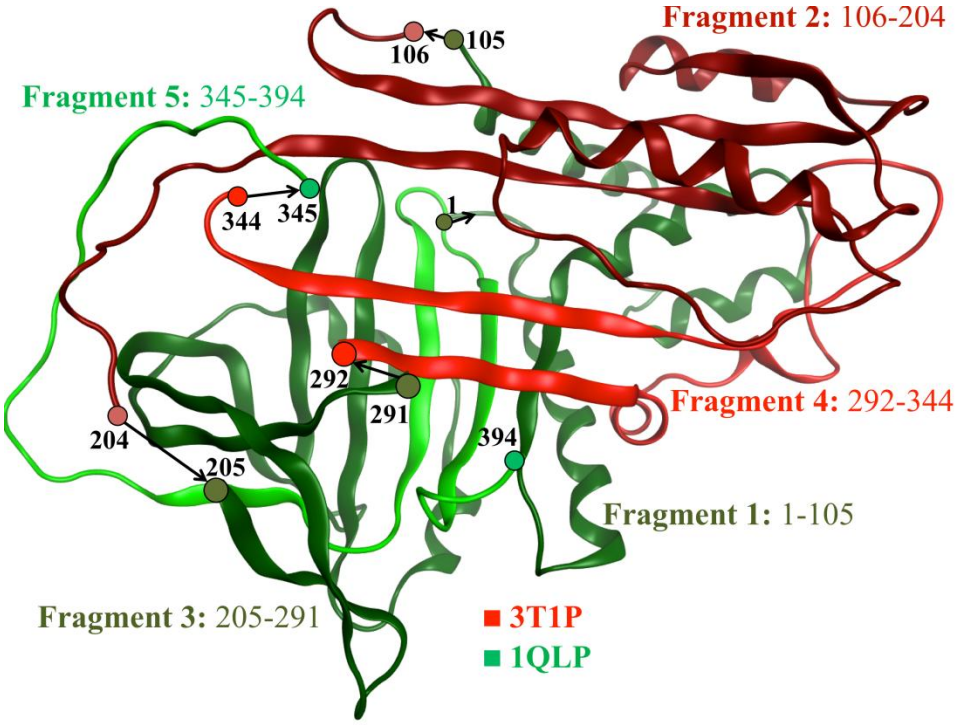


Figure 1.1.6. The fragments of structures 1QLP (green) and 3T1P (red) used to homology model the M* intermediate state of α1AT. β-sheet A is the red beta sheet across the top half of the model and β-sheet B is the green beta sheet across the bottom of the model. Residue numbers at the start and end of each fragment transition are labeled with an arrow in the N_{term} to C_{term} direction. Shades of green and red distinguish discontinuous fragments from the same initial crystal structure (light/dark green for 1QLP fragments and light/dark red for 3T1P fragments).

Analysis of α1AT Structures and their Potential Binding Sites

All of the 80 *in vitro* screened compounds, including S-(4-nitrobenzyl)-6-thioguanosine, were docked into every potential binding site to assess if the

computational result is comparable to the *in vitro* screening. A binding site able to dock S-(4-nitrobenzyl)-6-thioguanosine with a lower binding energy than the 79 other compounds would be a promising site for further experimental investigations.

Six putative binding sites were predicted among the three available protein models: M-α1AT, Z-α1AT and intermediate M* (Figure 1.1.7). SITE1 and SITE5 were both exclusively available in the M* model and are located in the RCL insertion site. SITE2 was found in all three models. It is also where the compound citrate, previously reported to lower polymerization rates[30] has been observed to bind in the 3CWM wild type structure. Also found in all three models are: SITE3, a large cavity adjacent to SITE2; SITE4 situated near the C_{term} edge of β-sheet A; and SITE6 located near the N_{term} edge of β-sheet A. SITE6 is partially occluded in the M* model due to the expansion of β-sheet A.

S-(4-nitrobenzyl)-6-thioguanosine Binds at the RCL Insertion Site or on the Edge of β-sheet A

Docking of all 80 small molecules was performed with each model and at each putative binding site in order to compare how strongly S-(4-nitrobenzyl)-6-thioguanosine binds relative to the 79 other experimentally tested compounds. These results are summarized in Table 1.1.1 and present two possible binding sites where S-(4-nitrobenzyl)-6-thioguanosine can favorably bind to block RCL insertion. Results from docking at SITE5, the RCL insertion site, show S-(4-nitrobenzyl)-6-thioguanosine ranking first among the other 79 ligands which may suggest a mechanism where the RCL is directly blocked at the RCL insertion site. Interestingly, S-(4-nitrobenzyl)-6-thioguanosine is also found to rank first in the wild type model when docked at SITE6. Figure 1.1.8 compares the location of SITE6 in both the M- and Z-α1AT structures which illustrates how binding of S-(4-nitrobenzyl)-6-thioguanosine at SITE6 may prevent the expansion of β-sheet A and possibly prevent RCL insertion. Lesser sites of interest, which only rank S-(4-nitrobenzyl)-6-thioguanosine in the top 10% of ligands, are SITES 1 and 2. SITE1 is also part of the RCL insertion site. SITE2 has been previously reported as the binding site for citrate which can also prevent polymerization and whose mechanism of action has yet to be determined.[30]

Residue Interactions with S-(4-nitrobenzyl)-6-thioguanosine

Nearby residues that interact with S-(4-nitrobenzyl)-6-thioguanosine at the top ranking sites (SITE1, SITE2, SITE5) from the M* intermediate model and the single site (SITE6) from the wild type model are described in Table 1.1.2. This information provides the basis for guiding further validations of these binding sites using techniques such as mutagenesis and molecular dynamics. Supplemental Figures 1.1.S1-1.S4 contain additional details about the type of interactions formed between individual atoms of S-(4-nitrobenzyl)-6-thioguanosine and the nearby residues listed in Table 1.1.2.

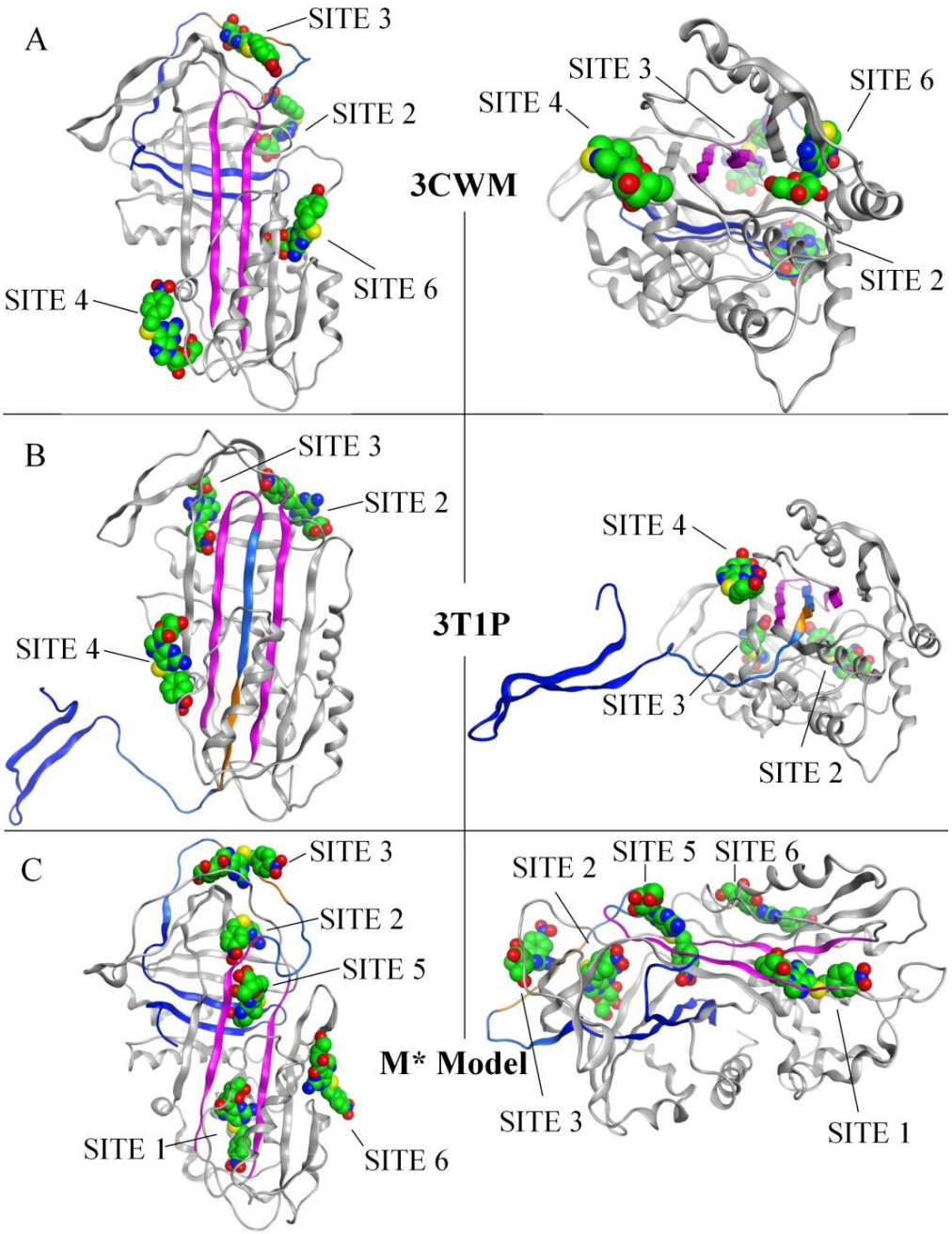


Figure 1.1.7. Binding Sites for S-(4-nitrobenzyl)-6-thioguanosine. Two protein ribbon models are shown for each structure: (A) 3CWM, (B) 3T1P and (C) M* Model. The left model and right representations in each panel are rotated 90° with respect to one another. The best binding poses for S-(4-nitrobenzyl)-6-thioguanosine at each available binding site are shown with space filling atoms with the carbon atoms colored green. (Purple) Strands 3 and 5 from β-sheet A. (Dark blue) C_{term} loop within β-sheet B. (Light blue) RCL. (Orange) Residues of the RCL corresponding to the analogous 6-mer peptide.

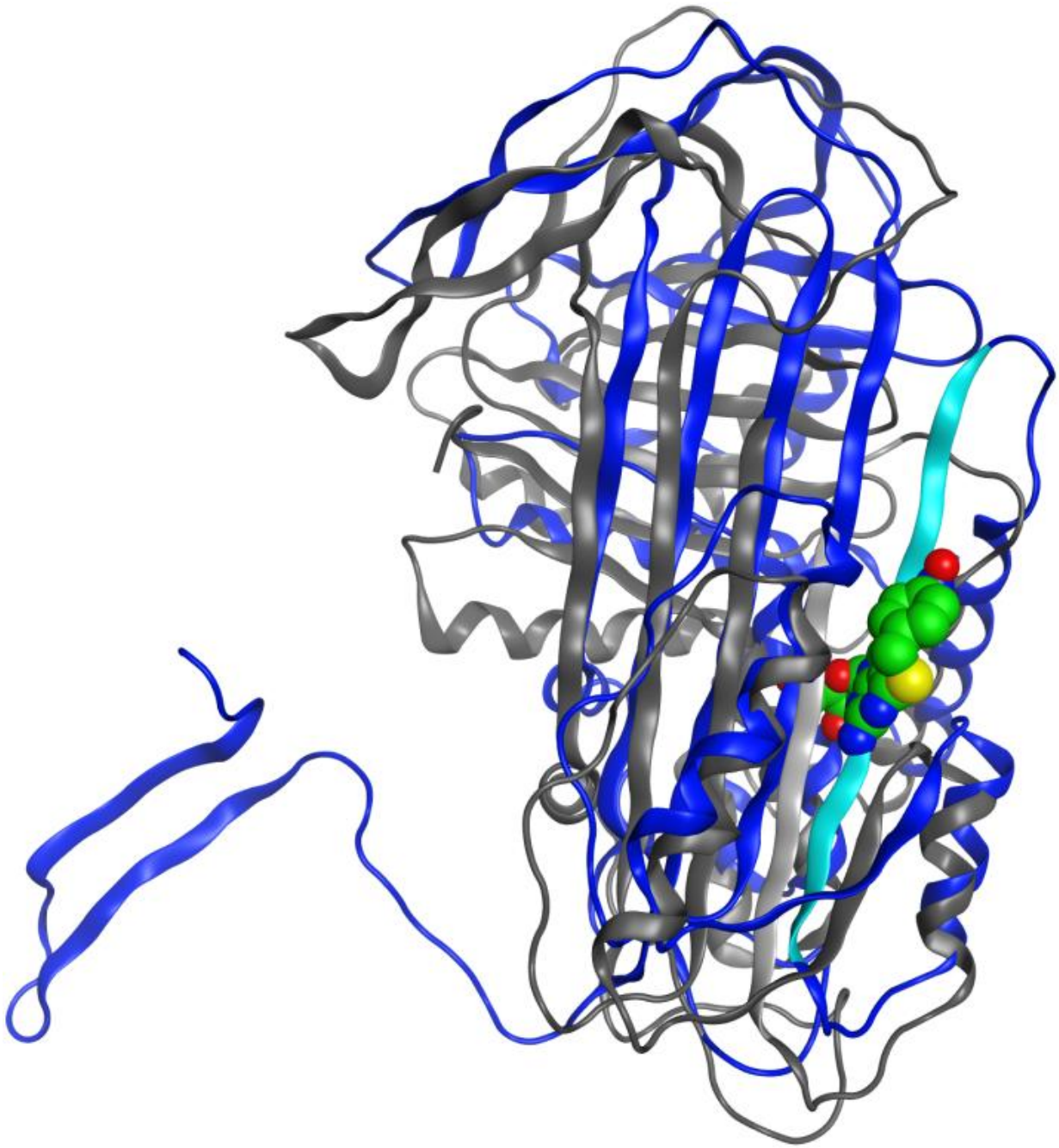


Figure 1.1.8. Two crystal structures for α1AT are superimposed and represented in ribbon diagram. S-(4-nitrobenzyl)-6-thioguanosine is represented with space filling atoms and positioned at SITE6 for the M-α1AT structure (1QLP). (Dark blue) Z-α1AT structure (3T1P) with an expanded β-sheet A. (Dark grey) Wild type structure 3CWM with β-sheet A not expanded into SITE6. (Light blue) expanded β-strand s2A in structure 3T1P, which occupies SITE6. (Light grey) β-strand s2A in structure 1QLP adjacent to SITE6.

Virtual Screening

Figure 1.1.9 shows the ranking of B9 in a virtual screening of the 80 LOPAC set and 1598 NCI Diversity set of compounds. The results of the screening place B9 in the top 1% of predicted binders which suggests that the screening model can work to find lead compounds. There were 16 compounds (unlisted at this time) binding better than B9 which are now prioritized to be experimental screened. Lead compounds found in the NCI Diversity dataset will be useful in producing several unique drug candidates to treat alpha-1-antitrypin deficiency. Identification of additional hit compounds from these leads will also offer support for improving the modeling and binding site location for future screening efforts.

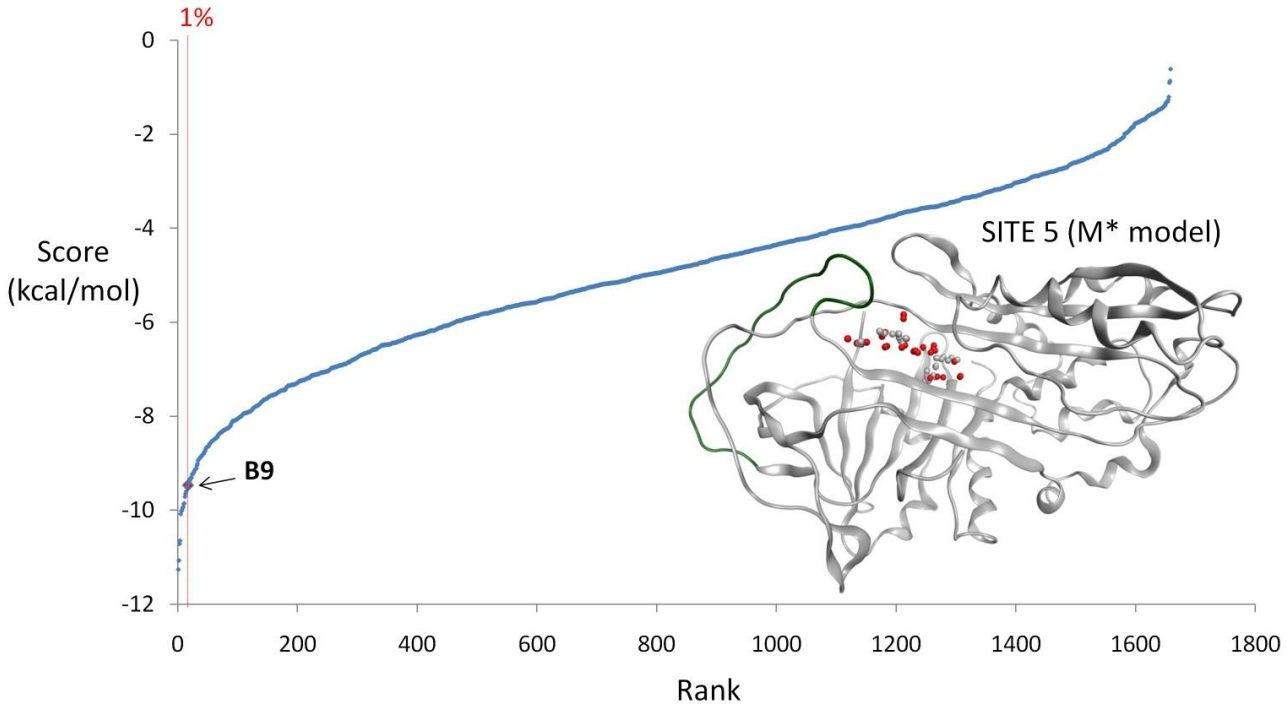


Figure 1.1.9. Virtual Screening. Shown is the virtual screening result of docking the NCI Diversity Set of 1598 compounds and 80 LOPAC Set of compounds into SITE 5 of the M* model. The M* model and SITE 5 are depicted in protein ribbon format with red and white spheres at SITE 5. The RCL is colored green. B9 is shown to rank within the top 1% of screening compounds.

Discussion

Currently, the only available and effective treatment to correct for the loss of α1AT function in α1ATD associated with liver disease is orthotropic liver transplantation.

For lung disease, augmentation therapy is the only specific regiment that is thought to slow down disease progression, although this still requires formal proof through well-controlled clinical trials.[36] As these treatments are expensive, labor intensive and associated with side effects, the need for novel treatments are indeed in high-demand. With Z-α1AT polymerization being responsible for the development of the disease, blocking its aggregation by small molecules [27,30] appears to be a promising strategy to cure Z-α1ATD.

Here we report an integrated *in vitro* and *in silico* approach which allows discovering and characterizing potent small molecules that disrupt the pathological polymerization of Z-α1AT. The *in vitro* microplate assay which enables the identification of small molecules able to block the insertion of a modified 6-mer peptide into the s4A cavity, provides quantitative data with reproducibility, sensitivity and rapid throughput. Our results validate the utility of the *in vitro* screening assay and identify S-(4-nitrobenzyl)-6-thioguanosine as inhibitor of Z-α1AT polymerization. With a molecular weight of 434.43 Da, 4 H-bond donors, 11 H-bond acceptors and a low lipophilicity coefficient (XLogP3 =1.1), this compound presents a strong drug-like profile according to the Lipinski rule of five.[37] From IC_{50} determination and structure-activity relationship studies, we also found one of its structural homologues which differs by a single amino group and does not prevent aggregation. This suggests that an interaction with the amino group may be important to counteract the insertion of the modified 6-mer peptide. The microplate assay has been designed to identify any inhibitor that can impede the insertion of the RCL into the s4A cavity, but it does not exclude the discovery of small molecules that can bind outside of the s4A cavity, causing a conformational rearrangement that still precludes RCL insertion. Molecular docking experiments were carried out to investigate binding of S-(4-nitrobenzyl)-6-thioguanosine at several potential binding locations, in addition to the s4A cavity. A chimera homology model of the intermediate state, M* model, was built in order to allow investigation of the s4A cavity. Until now, previous studies have only used molecular docking to investigate the binding of small molecules into experimentally resolved structures and only at sites other than the s4A cavity.[27]

The development for the first time of an atomistic M* model reveals a mechanism through which S-(4-nitrobenzyl)-6-thioguanosine may inhibit Z-α1AT polymerization by either competing with the RCL at the s4A insertion site (SITE5) or by blocking the s4A cavity from forming by binding at a nearby location (SITE6). To definitely discriminate between these two binding sites, additional structural studies will need to be carried out beyond the scope of the present work which will include selected mutagenesis and molecular dynamics (MD). MD simulations, which provide an ensemble of various conformations of M*, will account for the protein flexibility[38,39] and will aid in refining the docking results.

Table 1.1.1. Docking results from M-α1AT, Z-α1AT, and M* model with S-(4-nitrobenzyl)-6-thioguanosine and the 79 other small molecules.

Structure	SITE[α]	Lowest Energy[β] (kcal/mol)	B9 Energy[γ] (kcal/mol)	B9 Rank[δ]
M-α1AT (3CWM)	SITE1	n/a[Θ]	n/a	n/a
	SITE2	-7.0	-5.6	19th
	SITE3	-6.2	-5.3	8th
	SITE4	-6.4	-3.4	25th
	SITE5	n/a	n/a	n/a
	SITE6	-7.6	-7.6	1st
Z-α1AT (3T1P)	SITE1	n/a	n/a	n/a
	SITE2	-6.5	-3.9	34th
	SITE3	-9.0	-5.8	13th
	SITE4	-4.6	-3.6	19th
	SITE5	n/a	n/a	n/a
	SITE6	n/a	n/a	n/a
M* Model	SITE1	-10.7	-7.9	7th
	SITE2	-10.7	-8.8	5th
	SITE3	-5.1	-4.6	9th
	SITE4	n/a	n/a	n/a
	SITE 5	-8.4	-8.4	1st
	SITE6	-6.2	-4.22	12th

[α] Site number where S-(4-nitrobenzyl)-6-thioguanosine (B9) was docked. [β] Lowest observed binding energy (kcal/mol) for any of the 80 docked compounds. [γ] Predicted binding energy (kcal/mol) for B9. [δ] Rank of B9 relative to the binding energies for all 80 docked compounds. [Θ] Site numbers that are not found in a given model are noted by a not applicable symbol (n/a).

Table 1.1.2. Residues interacting with S-(4-nitrobenzyl)-6-thioguanosine in top scoring binding sites.

M* Model	SITE 1	S34, I35, A37, F38, L41, L149, T157, F159, A160, L161, V162, N163, Y164, L276, F289, L304, K305, L306, K308, A309, V310, H311
	SITE 2	W171, E172, R173, P174, F175, R200, M203, F204, N205, L218, M219, K220, Y221, F229, E256, D257, R258, L263, L265, I317, D318, F329, E331
	SITE 5	F28, K145, I146, I165, F166, F167, K168, V314, L315, C316, I 317, D318, E319, K320, G321, T322, E323, A324, M351, F361
M-α1AT (3CWM)	SITE 6	S56, T59, A60, M63, L100, N104, Q105, L112, T113, T114, G115, N116, G1117, Y138, H139, S140, E141, Y160, G164, N186, Y187, I188

Acknowledgement

The authors thank Professor Lomas, Cambridge Institute for Medical Research, University of Cambridge, UK, for his help in providing the purified human Z- and M-α1AT.

Funding Sources

This study was funded in part by the Alpha-1 Foundation and by the Physicians Medical and Education Research Foundation at the University of Tennessee Medical Center.

Abbreviations

α1AT, α1-antitrypsin; RCL, reactive center loop; bPEG-peptide, biotin-polyethylene glycol-peptide; Ab, antibody; B9, S-(4-nitrobenzyl)-6-thioguanosine; MW, molecular weight; C_{term}, C-terminal; N_{term}, N-terminal. RP-HPLC, reverse phase high-pressure liquid chromatography; MOE, molecular operating environment; PDB, protein database.

Supporting Information

Supplemental Figures 1.1.S1-1.S4 show details about the type of interactions formed between the individual atoms of S-(4-nitrobenzyl)-6-thioguanosine and nearby residues.

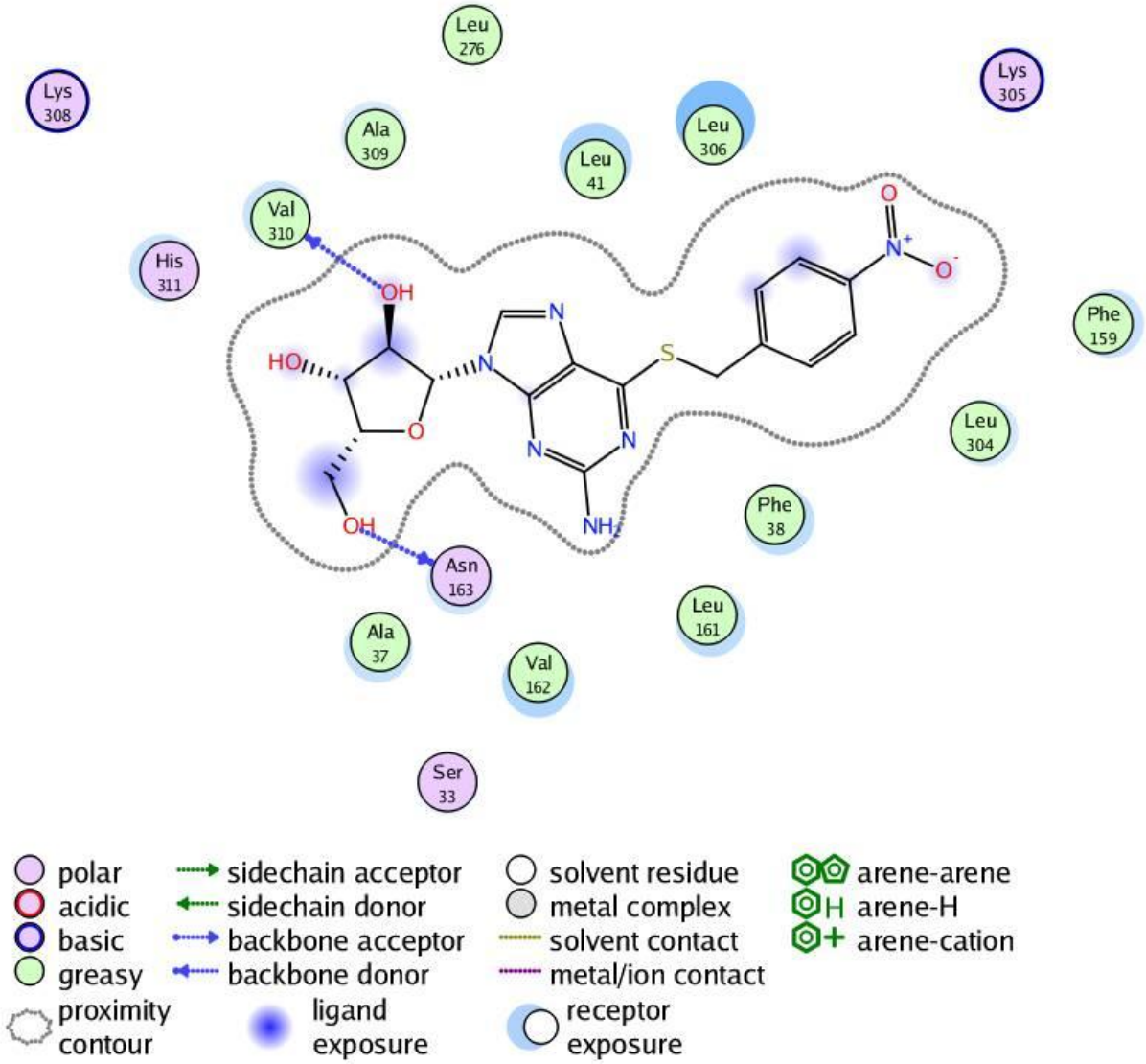


Figure 1.1.S1. 2-D contour and interaction map generated in MOE for S-(4-Nitrobenzyl)-6-thioguanosine at SITE1 in the M* intermediate state structure

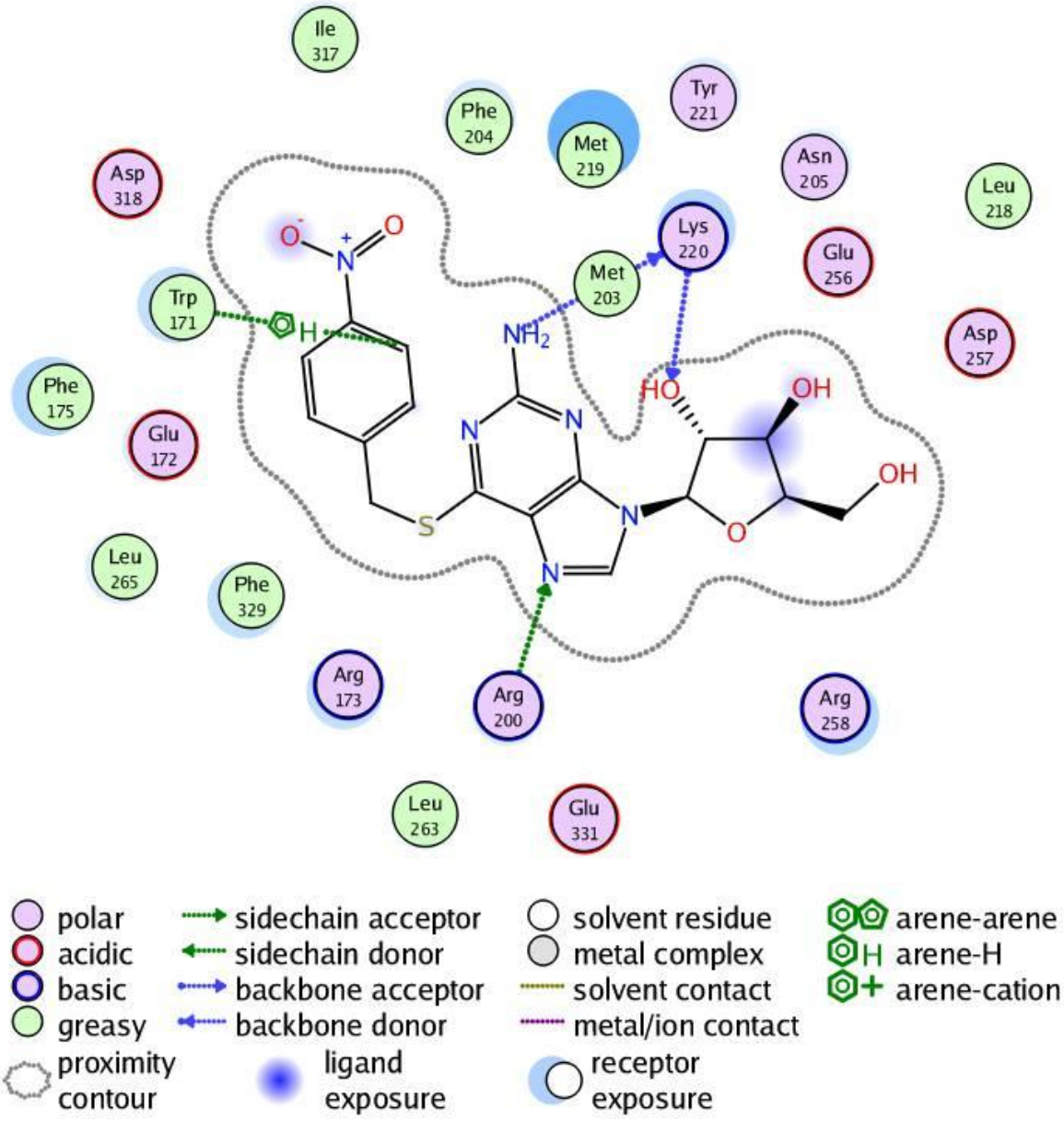


Figure 1.1.S2. 2-D contour and interaction map generated in MOE for for S-(4-Nitrobenzyl)-6-thioguanosine at SITE2 in the M* intermediate state structure

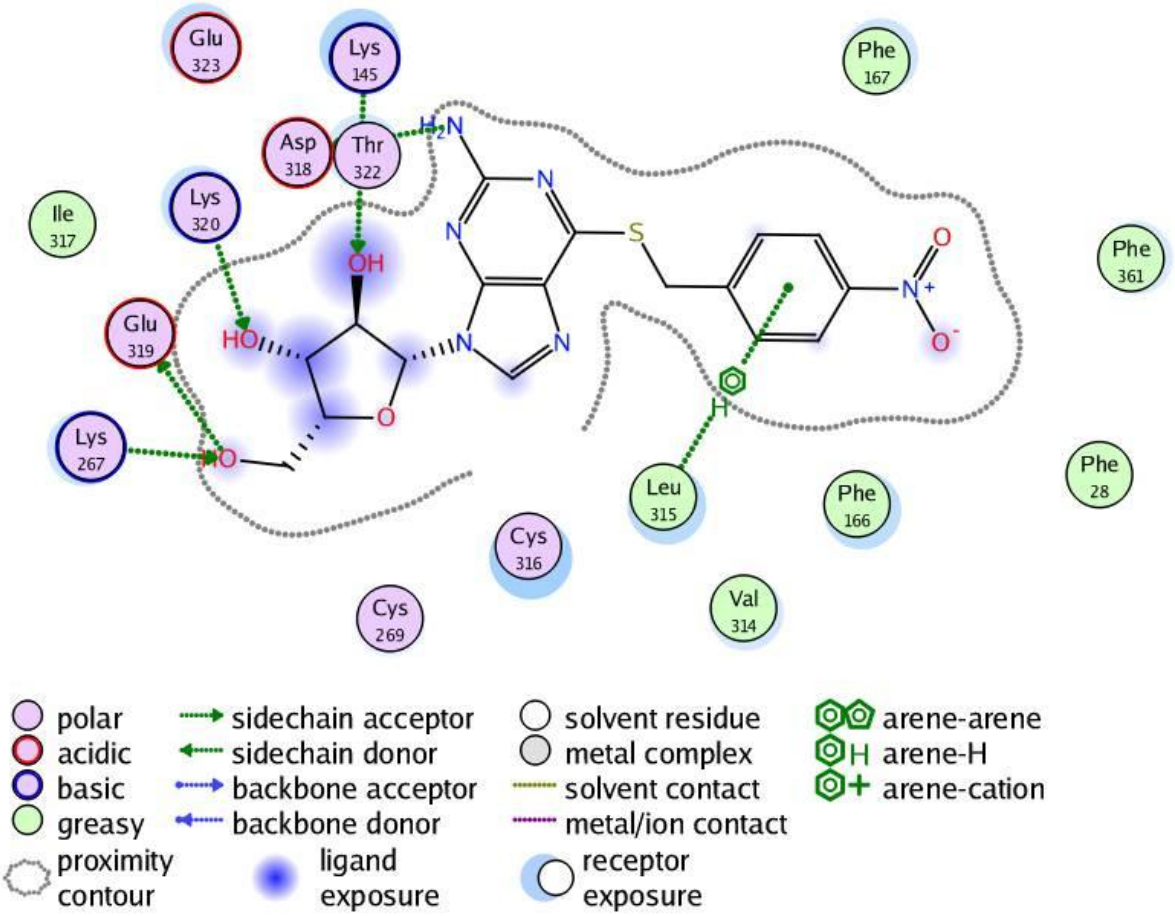


Figure 1.1.S3. 2-D contour and interaction map generated in MOE for S-(4-Nitrobenzyl)-6-thioguanosine at SITE5 in the M* intermediate state structure

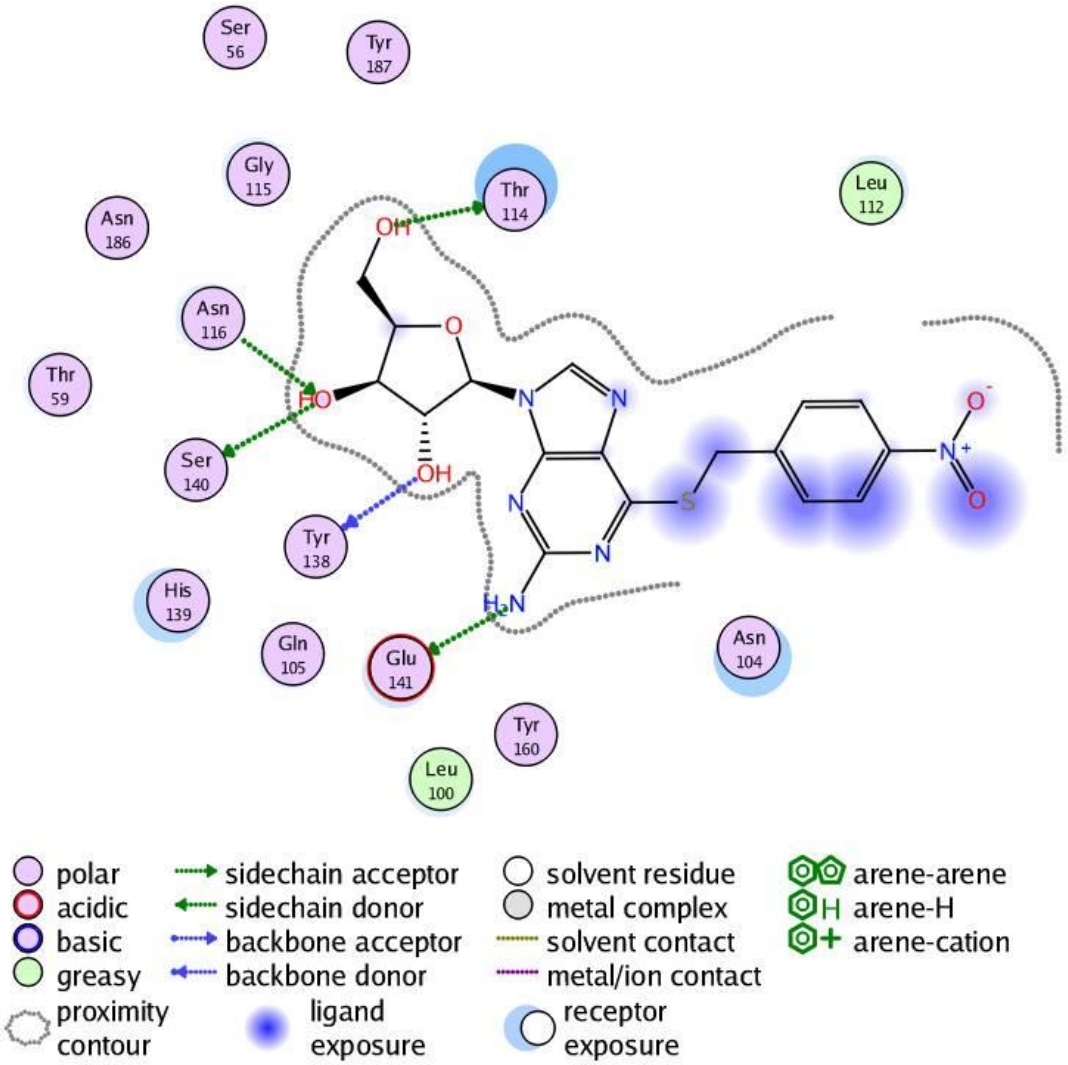


Figure 1.1.S4. 2-D contour and interaction map generated in MOE for S-(4-Nitrobenzyl)-6-thioguanosine at SITE6 in the 3CWM structure

References

(1) Brantly, M., Nukiwa, T., and Crystal, R. G. 1988. Molecular basis of alpha-1-antitrypsin deficiency. *Am. J. Med. 84*, 13–31.

(2) Gooptu, B., Dickens, J. A., and Lomas, D. A. 2014. The molecular and cellular pathology of α1-antitrypsin deficiency. *Trends Mol. Med. 20*, 116–27.

(3) Perlmutter, D. H. 2011. Alpha-1-antitrypsin deficiency: importance of proteasomal and autophagic degradative pathways in disposal of liver disease-associated protein aggregates. *Annu. Rev. Med. 62*, 333–45.

(4) Huntington, J. A., Read, R. J., and Carrell, R. W. 2000. Structure of a serpin-protease complex shows inhibition by deformation. *Nature; Nat. 407*, 923–926.

(5) Whisstock, J. C., and Bottomley, S. P. 2006. Molecular gymnastics: serpin structure, folding and misfolding. *Curr. Opin. Struct. Biol. 16*, 761–768.

(6) Dafforn, T. R., Mahadeva, R., Elliott, P. R., Sivasothy, P., and Lomas, D. A. 1999. A kinetic mechanism for the polymerization of alpha1-antitrypsin. *J. Biol. Chem. 274*, 9548–9555.

(7) Lomas, D. A. 2000. Loop-sheet polymerization: the mechanism of alpha1-antitrypsin deficiency. *Respir. Med. 94 Suppl C*, S3–6.

(8) De Serres, F. J., and Blanco, I. 2012. Prevalence of α1-antitrypsin deficiency alleles PI*S and PI*Z worldwide and effective screening for each of the five phenotypic classes PI*MS, PI*MZ, PI*SS, PI*SZ, and PI*ZZ: a comprehensive review. *Ther. Adv. Respir. Dis. 6*, 277–95.

(9) Eriksson, S., Carlson, J., and Velez, R. 1986. Risk of cirrhosis and primary liver cancer in alpha 1-antitrypsin deficiency. *New Engl. J. Med. New Engl. J. Med. 314*, 736–739.

(10) An, J. K., Blomenkamp, K., Lindblad, D., and Teckman, J. H. 2005. Quantitative isolation of alphalAT mutant Z protein polymers from human and mouse livers and the effect of heat. *Hepatol. (Baltimore, Md.); Hepatol. (Baltimore, Md.) 41*, 160–167.

(11) Hussain, M., Mieli-Vergani, G., and Mowat, A. P. 1991. Alpha 1-antitrypsin deficiency and liver disease: clinical presentation, diagnosis and treatment. *J. Inherit. Metab. Dis. J. Inherit. Metab. Dis. 14*, 497–511.

(12) Ogushi, F., Fells, G. A., Hubbard, R. C., Straus, S. D., and Crystal, R. G. 1987. Z-type alpha 1-antitrypsin is less competent than M1-type alpha 1-antitrypsin as an inhibitor of neutrophil elastase. *J. Clin. Invest. 80*, 1366–74.

(13) Lomas, D. A., Evans, D. L., Stone, S. R., Chang, W. S., and Carrell, R. W. 1993. Effect of the Z mutation on the physical and inhibitory properties of alpha 1-antitrypsin. *Biochem. (John Wiley Sons); Biochem. 32*, 500–508.

(14) Elliott, P. R., Bilton, D., and Lomas, D. A. 1998. Lung polymers in Z alpha(1)-antitrypsin deficiency-related emphysema. *Am. J. Respir. Cell Mol. Biol. 18*, 670–674.

(15) Mahadeva, R., Atkinson, C., Li, Z., Stewart, S., Janciauskiene, S., Kelley, D. G., Parmar, J., Pitman, R., Shapiro, S. D., and Lomas, D. A. (2005) Polymers of Z alpha1-antitrypsin co-localize with neutrophils in emphysematous alveoli and are chemotactic in vivo. *Am. J. Pathol. 166*, 377–386.

(16) Parmar, J. S., Mahadeva, R., Reed, B. J., Farahi, N., Cadwallader, K. A., Keogan, M. T., Bilton, D., Chilvers, E. R., and Lomas, D. A. 2002. Polymers of alpha(1)-antitrypsin are chemotactic for human neutrophils: a new paradigm for the pathogenesis of emphysema. *Am. J. Respir. Cell Mol. Biol. 26*, 723–730.

(17) Lomas, D. A., Perlmutter, D. H., and Uversky, V. N. 2010. Alpha-1Antitrypsin Deficiency, in *Protein Misfolding Diseases - Current and Emerging Principles and Therapies* (Ramirez-Alvarado, M., Kelly, J. W., and Dobson, C. M., Eds.), pp 403–424. A John Wiley & Sons, Inc., Hoboken, New Jersey.

(18) Mahadeva, R., Dafforn, T. R., Carrell, R. W., and Lomas, D. A. 2002. 6-mer peptide selectively anneals to a pathogenic serpin conformation and blocks polymerization. Implications for the prevention of Z alpha(1)-antitrypsin-related cirrhosis. *J. Biol. Chem. 277*, 6771–6774.

(19) Gooptu, B., Hazes, B., Chang, W. S., Dafforn, T. R., Carrell, R. W., Read, R. J., and Lomas, D. A. 2000. Inactive conformation of the serpin alpha(1)-antichymotrypsin indicates two-stage insertion of the reactive loop: implications for inhibitory function and conformational disease. *Proc. Natl. Acad. Sci. U. S. A. 97*, 67–72.

(20) Purkayastha, P., Klemke, J. W., Lavender, S., Oyola, R., Cooperman, B. S., and Gai, F. 2005. Alpha 1-antitrypsin polymerization: a fluorescence correlation spectroscopic study. *Biochemistry 44*, 2642–2649.

(21) Ekeowa, U. I., Freeke, J., Miranda, E., Gooptu, B., Bush, M. F., Perez, J., Teckman, J., Robinson, C. V, and Lomas, D. A. 2010. Defining the mechanism of polymerization in the serpinopathies. *Proc. Natl. Acad. Sci. U. S. A. 107*, 17146–17151.

(22) Yamasaki, M., Li, W., Johnson, D. J., and Huntington, J. A. 2008. Crystal structure of a stable dimer reveals the molecular basis of serpin polymerization. *Nature 455*, 1255–1258.

(23) Yamasaki, M., Sendall, T. J., Pearce, M. C., Whisstock, J. C., and Huntington, J. A. 2011. Molecular basis of alpha1-antitrypsin deficiency revealed by the structure of a domain-swapped trimer. *EMBO Rep. 12*, 1011–1017.

(24) Burrows, J. A., Willis, L. K., and Perlmutter, D. H. 2000. Chemical chaperones mediate increased secretion of mutant alpha 1-antitrypsin (alpha 1-AT) Z: A potential pharmacological strategy for prevention of liver injury and emphysema in alpha 1-AT deficiency. *Proc. Natl. Acad. Sci. U. S. A. 97*, 1796–1801.

(25) Devlin, G. L., Parfrey, H., Tew, D. J., Lomas, D. A., and Bottomley, S. P. 2001. Prevention of polymerization of M and Z alpha1-Antitrypsin (alpha1-AT) with trimethylamine N-oxide. Implications for the treatment of alpha1-at deficiency. *Am. J. Respir. Cell Mol. Biol. 24*, 727–732.

(26) Teckman, J. H. 2004. Lack of effect of oral 4-phenylbutyrate on serum alpha-1-antitrypsin in patients with alpha-1-antitrypsin deficiency: a preliminary study. *J. Pediatr. Gastroenterol. Nutr. 39*, 34–37.

(27) Mallya, M., Phillips, R. L., Saldanha, S. A., Gooptu, B., Leigh, S. C., Termine, D. J., Shirvani, A. M., Wu, Y., Sifers, R. N., Lomas, D. A., Brown, S. C., and Abagyan, R. 2007. Small molecules block the polymerization of Z alpha1-antitrypsin and increase the clearance of intracellular aggregates. *J. Med. Chem. 50*, 5357–5363.

(28) Parfrey, H., Mahadeva, R., Ravenhill, N. A., Zhou, A., Dafforn, T. R., Foreman, R. C., and Lomas, D. A. 2003. Targeting a surface cavity of alpha 1-antitrypsin to prevent conformational disease. *J. Biol. Chem. 278*, 33060–33066.

(29) Makarananda, K., and Neal, G. E. 1992. Competitive ELISA. *Methods Mol. Biol. 10*, 267–72.

(30) Pearce, M. C., Morton, C. J., Feil, S. C., Hansen, G., Adams, J. J., Parker, M. W., and Bottomley, S. P. 2008. Preventing serpin aggregation: The molecular mechanism of citrate action upon antitrypsin unfolding. *Protein Sci. 17*, 2127–2133.

(31) Elliott, P. R., Pei, X. Y., Dafforn, T. R., and Lomas, D. A. 2000. Topography of a 2.0 A structure of alpha1-antitrypsin reveals targets for rational drug design to prevent conformational disease. *Protein Sci. a Publ. Protein Soc. 9*, 1274–1281.

(32) 2012. Chemical Computing Group - Citing MOE.

(33) Labute, P. 2009. Protonate3D: Assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins Struct. Funct. Bioinforma. 75*, 187–205.

(34) Halgren, T. A. 1999. OBForceFieldMMFF94. J. Comput. Chem. 20, 720-729. No Title.

(35) Labute, P. 2008. The generalized Born/volume integral implicit solvent model: Estimation of the free energy of hydration using London dispersion instead of atomic surface area. *J. Comput. Chem. 29*, 1693–1698.

(36) Gøtzsche, P. C., and Johansen, H. K. 2010. Intravenous alpha-1 antitrypsin augmentation therapy for treating patients with alpha-1 antitrypsin deficiency and lung disease. *Cochrane database Syst. Rev.* CD007851.

(37) Leeson, P. 2012. Drug discovery: Chemical beauty contest. *Nature 481*, 455–6.

(38) Cheng, L. S., Amaro, R. E., Xu, D., Li, W. W., Arzberger, P. W., and McCammon, J. A. 2008. Ensemble-Based Virtual Screening Reveals Potential Novel Antiviral Compounds for Avian Influenza Neuraminidase. *J. Med. Chem. 51*, 3878–3894.

(39) Demir, Ö., Baronio, R., Salehi, F., Wassman, C.D., Hall, L., Hatfield, G.W., Chamberlin, R., Kaiser, P., Lathrop, R.H., Amaro, R.E. 2011. Ensemble-based computational approach discriminates functional activity of p53 cancer and rescue mutants. *PLoS Comp*.

CHAPTER 2.1
A COMPUTATIONAL APPROACH PREDICTING CYP450 METABOLISM AND ESTROGENIC ACTIVITY OF AN ENDOCRINE DISRUPTING COMPOUND (PCB-30)

A version of this chapter was originally published by Jason B. Harris and Melanie L. Eldridge, Gary Sayler, Fu-Min Menn, Alice C. Layton, Jerome Baudry:

Jason B. Harris and Melanie L. Eldridge, Gary Sayler, Fu-Min Menn, Alice C. Layton, Jerome Baudry. "A Computational Approach Predicting CYP450 Metabolism and Estrogenic activity of an Endocrine Disrupting Compound (PCB-30)". *Environmental Toxicology & Chemistry* (2014). In Press. DOI: 10.1002/etc.2595

The work and writing in this article was equally contributed to by Jason B. Harris and Melanie L. Eldridge. M.L. Eldridge designed the estrogen assay protocol. J.B. Harris designed all computational models. J.B. Harris and M.L Eldridge co-designed the P450 assay protocol. J.B. Harris carried out experiments for the P450 assay. J.B. Harris and M.L. Eldridge each carried out experiments for the estrogen assay. Fu-Min Menn carried out GC/MS analysis on metabolite extracts. Gary Sayler, Alice C. Layton, and Jerome Baudry served as faculty mentors and assisted in manuscript revisions.

Abstract

Endocrine disrupting chemicals (EDCs) influence growth and development through interactions with the hormone system, often through binding to hormone receptors such as the estrogen receptor. Computational methods can predict EDC activity of unmodified compounds, but approaches predicting activity following metabolism are lacking. This study uses a well-known environmental contaminant, PCB-30 (2,4,6-trichlorobiphenyl), as a prototype EDC and integrates predictive (computational) and experimental methods to determine its metabolic transformation by CYP3A4 and CYP2D6 into estrogenic byproducts. Computational predictions suggest that hydroxylation of PCB-30 occurs at the 3- or 4-phenol positions and leads to metabolites that bind more strongly than the parent molecule to the human estrogen receptor alpha (hER-α). GC/MS experiments confirmed that the primary metabolite for CYP3A4 and CYP2D6 is 4-hydroxy-PCB-30, and the secondary metabolite is 3-hydroxy-PCB-30. Cell-based bioassays (bioluminescent yeast expressing hER-α) confirmed that hydroxylated metabolites are more estrogenic than PCB-30. These experimental results support the applied model's ability to predict the metabolic and estrogenic fate of PCB-30, which could be used to identify other EDCs involved in similar pathways.

Introduction

Endocrine disrupting chemicals (EDCs) can influence growth and developmental processes in humans and animals[1,2] which has led to a sustained effort toward identifying and characterizing potential endocrine disrupting compounds[3-5] including polychlorinated biphenyls (PCBs). PCBs are environmentally widespread and recalcitrant pollutants with multiple mechanisms for endocrine disruption, including interfering with estrogen hormone signaling[6-10] and are of particular concern due to their bioaccumulation through the food chain to relatively high concentrations within the adipose tissues of top predators.[11-13] One of the most estrogenic metabolites of all PCBs is 4-hydroxy-PCB-30.[14] Its parent molecule, PCB-30, is one of 209

polychlorinated biphenyl (PCB) congeners which has been used as a prototypic molecule for many other metabolite and estrogenic studies.[15,16,3]

Cytochrome P450s are primarily responsible for metabolizing PCBs and other EDCs, most often by adding hydroxyl groups to their aromatic rings. Hydroxylation serves to make exogenous compounds more polar and thereby more easily expelled from cells through active transport mechanisms.[17] In many cases, hydroxylation of a compound gives the corresponding metabolites a higher binding affinity to estrogen or other hormone receptors than the parent molecule.[18-24] This process is known as bioactivation and experimental methods able to detect it are prohibitively costly and time consuming. Predictive approaches are needed in order to screen a large number of chemicals that may have EDC properties or obtain them through P450 metabolism.[1,5] Ryberg[25] has produced a program dubbed SMARTCyp which pre-calculates reactivity rules for atoms in P450 active sites and uses 2D structural similarity rules to assign reactivity scores to new molecules. Virtual docking has been shown to provide useful *de novo* predictions about compound accessibility and binding affinity to a target protein which is directly relatable bioactivity. Virtual docking has for instance been successfully used to identify active EDCs toward hormone receptor target [26,27] and CYP450 ligands[28-30]. Ginex[31] and colleagues presented recent work where the primary P450 metabolites of a compound were predicted using a computational 2D method, similar to SMARTCyp, and followed by bioactivity predictions using virtual docking to the androgen receptor.

In this work, an integrated computational and experimental approach (outlined in Figure 2.1.1) has been developed to identify chemicals that exhibit EDC properties following metabolic processing by specific cytochrome P450 enzymes. This approach is currently modeled with PCB-30 as a prototype molecule and applied to the estrogen receptor and specific P450s 3A4/2D6. Our model differs most notably from Ginex's scheme by including the use of docking in 3D P450 structures as a means to better assess atom accessibilities, a point noted by Ryberg as a limiting factor in 2D predictive methods.

Predicting the P450 metabolites of a ligand requires knowledge about i) the reactivity for each ligand atom and ii) the accessibility for each ligand atom within the active site of specific P450 enzymes. In the outlined model (Figure 2.1.1), SMARTCyp[25] is used to predict the relative reactivity for each ligand atom, with respect to specific P450 enzymes, and molecular (virtual) docking is used to predict the geometry (accessibility) of a ligand atom within a specific P450's active site. Any predicted P450 metabolites are then docked into the ligand binding domain of the human estrogen receptor alpha (hER-α) to assess binding affinity to this nuclear hormone receptor, and hence gauge any potential estrogenicity.

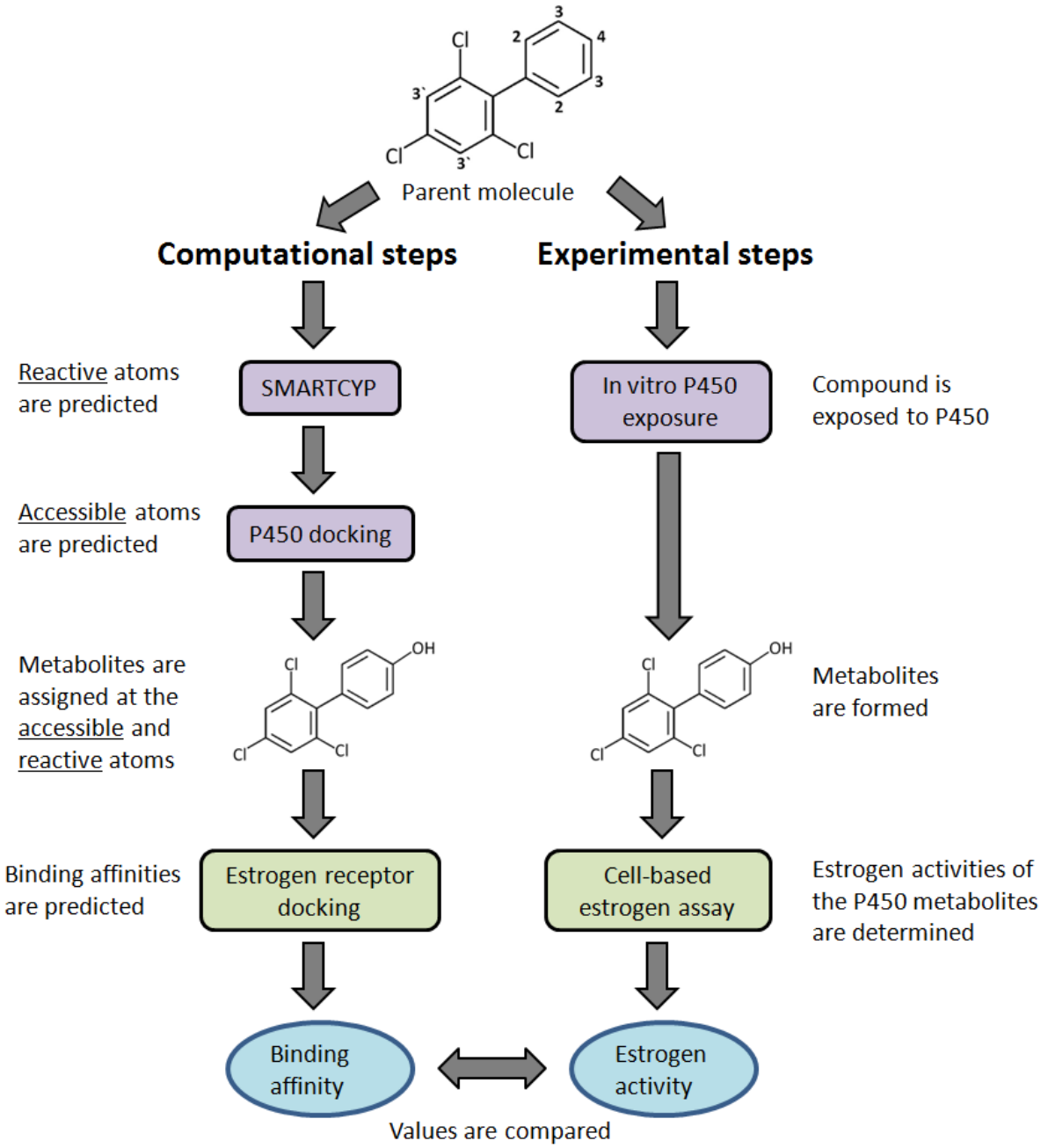


Figure 2.1.1. Computational and Experimental Approaches. Potentially reactive carbon atoms on the parent molecule are labeled by atom type (4, 3, 3', 2). (Left Side) Computational steps; (Right Side) Experimental steps. Comparable methods between the left and right sides of the diagram are coordinated by color (Purple, Blue, Green).

Metabolite predictions for PCB-30 were based on two specific CYP450 isoforms, CYP2D6 and CYP3A4, which together metabolize ~50% of known drugs[32,33] including other estrogen-like compounds.[34,35] CYP3A4 is known to metabolize a diverse range of compounds due to its large and flexible ligand binding site.[36,37] These two specific CYP450s are both involved in converting a notable EDC, tamoxifen, to more highly estrogenic 3- and 4-hydroxy species.[38] Part of the metabolic process involves the hydroxylation of a resonant ring, a functional group shared by both tamoxifen and PCB-30.

Computational predictions were validated using a series of complementary experimental techniques (Figure 2.1.1, right column). First, microsomal reaction mixtures (MRMs), obtained by exposing PCB-30 to human liver microsomes, were used to generate hydroxylated metabolites. The hydroxylated products in the MRMs were then structurally characterized using GC/MS and also functionally (i.e., assessing their estrogenic activities) in a recombinant bioluminescent yeast assay (BLYES)[39,40] which measures binding of chemicals to hER-α. The BLYES bioassay consists of *Saccharomyces cerevisiae* with the human estrogen receptor (alpha form) expressed continually from its genome as well as plasmid-based *lux* operon (*luxCDABE*) bioreporter genes under the control of human estrogen response elements.[40] The engineered yeast strain produces bioluminescence when exposed to estrogenic compounds and does not need exogenous substrates. These assays have been used extensively to measure endocrine responses to polychlorinated biphenyls (PCBs) and hydroxylated derivatives[41-43], polynuclear aromatic hydrocarbons (PAHs)[44], pesticides[45] and other compounds[43] as well as detection of estrogens in natural samples[41,46-49].

Using the integrated approaches outlined in Figure 2.1.1, two specific P450 enzymes (CYP2D6 and CYP3A4) are identified for the first time as responsible for producing the potent estrogenically active metabolites of PCB-30. Detailed biochemical predictions observed as a part of these findings indicate a potential to use this modeling approach to screen additional compounds through multi-protein biochemical pathways. A future full-scale use of this method could screen large databases of chemicals against many different protein targets. Such a massive virtual screening approach is becoming more feasible as recent technology improvements are able to better scale docking simulations on supercomputers, allowing millions of compounds to be screened against multiple protein targets.[50] Complex system-level predictions for toxicity and therapeutics will continue to develop over the next decade as simulation technologies and computational power continue to improve,[51] and this work offers a contribution toward moving the field of molecular modeling in this forward direction of predicting biochemical pathway interactions.

Materials and Methods

Ligand Preparation

The structure for PCB-30 was obtained from the PubChem database (Compound CID: 37247). Metabolite structures were built using the program MOE version 2010[52]

and the PCB-30 parent molecule as a template. Hydrogen atoms were added using the Protonate-3D[53] facility in MOE with a pH condition of 7. Inclusion of pH conditions in determining protonation states is important since functional groups may be either ionized or neutral, affecting a molecules predicted bioactivity during docking simulations.[54] All structures were energy-minimized to a 0.05 kcal/mol·Å RMS (root mean square) energy gradient using the MMFF94s force field (Merck Molecular Force Field 94)[55] as implemented in MOE.

P450 Docking

To represent the active iron-oxygen intermediate state of the P450 heme, the crystal structures for cytochrome P450s CYP3A4 (PDB: 1TQN)[56] and CYP2D6 (PDB: 2F9Q)[57] were used following the protocol described in Baudry *et al.*, 2003.[30] Docking calculations were performed using MOE**.** Water molecules were deleted from the crystal structures. Hydrogen atoms and charges were assigned using Protonate-3D facility in MOE (pH 7) which estimates the individual pKa of every residue. The Site Finder facility in MOE was used to identify binding site locations. The default placement method, Triangle Matcher (matching 3 atoms at a time), was used to quickly sample 3000 ligand poses. Ligand bonds were allowed to be rotatable during placement. The returned poses were scored by their approximate free energies of binding using the Affinty dG[52] scoring method which estimates the binding free energy of a ligand in a given conformation inside the protein's active site. The top 100 scoring poses were retained for further calculations. Duplicate orientations were removed from these 100 returned poses. The ligand atoms and binding pocket residues (6 Å away) were refined with forcefield energy minimization (0.01kcal/mol·Å RMS energy gradient) using the MMFF94s forcefield. Side chains atoms were allowed to be partially flexible (6 Å away from the ligand pocket atoms) with a tethering weight set to 10kcal/mol/A^2. Final binding energy calculations for each of the remaining poses were calculated using the London dG[58] scoring method.

Estrogen Receptor Docking

The crystal structure for hER-α protein (PDB: 1L2I)[59] was used to dock PCB-30 and its predicted metabolites because it represents the agonist form of the receptor at a resolution of less than 2 angstroms. The docking protocol was developed using a test set of 2636 compounds from the DUD-ER[60] (Directory of Useful Decoys- Estrogen Receptor) database containing 67 known active molecules (i.e. ligands) and 2569 known decoys. Supplemental Figure 2.1.S1 shows that docking statistically distinguishes active from inactive compounds. Most of the active compounds could be recovered in a small fraction of this database after prioritization with docking scores. For example, 10% of the scored database was enriched to contain 85% of the original actives. Experimental testing of this top 10% would only include 8.1% of the original decoys (false positives), exhibiting significant enrichment over the database ratio of 38 decoys for every 1 active molecule. In the docking protocol, all atoms were removed except for chain A residues and the associated water molecules. Water molecule 10 is known to coordinate interactions within a hydrogen bond network involving GLU 353, ARG 394 and the binding of various hydroxylated ligands.[61, 62] Hydrogen atoms were added to the protein using the Protonate-3D facility in MOE (pH 7). The Site Finder

utility in MOE was used to identify the binding site in the structure of hER-α. For docking calculations, the Triangle Matcher placement method was used for initial ligand positioning. This was set to return 1000 poses. Ligand bonds were allowed to be rotatable during placement. MOE's internal scoring function for the placement of each conformer (E_place) was used to determine the best initial ligand pose, based on geometric fit. The ligand and binding pocket atoms for the best placed pose (within 6 Å from the ligand) were refined through energy minimization to a 0.01kcal/mol·Å RMS energy gradient using the MMFF94s forcefield and a tethering weight set to 10kcal/mol/A^2. A dielectric constant of 1 was used in the forcefield potential setup. Final energy scores were calculated using MOE's default scoring method, London dG.

SMARTCyp

The SMARTCyp web server (version 2.2) was used to assess the reactivity of individual atoms on PCB-30. SMARTCyp bases its reactivity predictions on a set of pre-calculated activity rules for chemical functional groups. The specific transformations that occur at a given atomic location are assigned from a list of the most commonly known transformations observed for particular functional groups and P450s.[25] More direct approaches, up to and including quantum approaches, can be used to more accurately evaluate reaction products than statistical methods, although such calculations are extremely time consuming and are not adequate for use in screening models.[63] In the case of PCB-30, the only functional group to consider is a phenolic ring for which hydroxylation is the most common transformation known to be catalyzed. Other transformations could in theory be assigned for molecules with different functional groups, though none are included in the scope of this current work. SMARTCyp also estimates an accessibility score for each possibly reactive atom to account for the size and shape of a specific P450's active site. For CYP3A4's active site, which is larger than that of CYP2D6 and allows free molecular rotation, atoms farthest from a molecule's center are considered to be more accessible. In the case of CYP2D6, atoms on the farthest ends of a molecule are treated as more accessible since CYP2D6's active site is narrower than that of CYP3A4, not allowing the molecules to freely rotate. Benchmarks for SMARTCyp accuracy are reported to be 91% for CYP2D6 and 76% for CYP3A4 with regards to identifying the top two sites of metabolism in sample sets of compounds.[25,64] The lower accuracy rate for CYP3A4 predictions is due to the larger and less restrictive binding cavity.

Yeast Bioassays

Strains *Saccharomyces cerevisiae* BLYES and BLYR[43] were grown in Yeast Minimal Media (YMM) without leucine and uracil overnight at 30^{o}C and with shaking to an OD_{600} of 1.0. All chemicals for the assay were initially suspended in 4% DMSO at a concentration of 10 mM. The standard (17β-estradiol) was serially diluted to generate a concentration range from 0.2 µM to 0.5 pM, while 4-hydroxy-PCB-30 was diluted from 2 µM to 5 pM and PCB-30 was diluted from 2 mM to 5 nM. Then 100 µl of each dilution (standards and samples) were spotted into the wells of 96-well plates. A 100-µl aliquot of culture was placed into each well of the 96-well plate, generating a concentration range from 0.1 µM to 0.25 pM for 17β-estradiol, 1 µM to 2.5 pM for 4-hydroxy-PCB-30, and 1 mM to 2.5 nM for PCB-30. For each test assay, plates were made in duplicate

such that one strain was added per plate (BLYR and BLYES), including controls. Bioluminescence was measured every 60 min for 6 h in a Perkin-Elmer Victor2 Multilabel Counter with an integration time of 1 s/well. Negative controls included wells with (i) medium + cells and (ii) medium + cells + DMSO. Log bioluminescence (counts per second) versus log sample dilution or chemical control concentration (M) were plotted, generating a sigmoidal curve for hormonally active compounds. A 50% effective concentration (EC_{50}) value was determined from the midpoint of the linear portion of the sigmoidal curve. The mean and standard deviation values were calculated from replicate EC_{50} values for standards to determine the variability between assays. The detection limit was determined by calculating the concentration of chemical at background bioluminescence plus three standard deviations of bioluminescence.

Microsomal Reaction Mixture

A stock solution ($1x10^{-2}$ M) of each chemical was prepared in DMSO and incubated with microsome mixtures according to the manufacturer's protocol. Three types of microsomal mixtures were used: a pool of microsomal extracts from 50 pooled human donor livers, microsomal extracts that were enriched for CYP2D6, and microsomal extracts that were enriched for CYP3A4 (Life Technologies, Grand Island NY). Pooled human microsomes contained the cytochromes CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP2E1, and CYP3A4. Briefly, 2 µl of 20 mM NADPH, 5 µl of microsome solution, and 183 µl of phosphate buffer were mixed with 2 µl of chemical stock solution in DMSO and incubated for 5 minutes with shaking at 37^{o}C. NADPH (10 µl of 20 mM) was added and the reaction mixture was incubated at 37^{o}C with shaking for 24 hours. The reaction was stopped by the addition of 200 µl of DMSO, yielding a final concentration of test chemical $1x10^{-4}$ M in 50% DMSO. Test chemical-microsome mixtures were then diluted and processed in the yeast assays as described in the methods section of Sanseverino et al. 2009. Three types of microsome mixtures were tested with both PCBs: 1) a pool of human liver microsomes from 50 people, 2) a pool of human liver microsomes with high CYP2D6 activity, and 3) a pool of human liver microsomes with high CYP3A4 activity.

Gas Chromatography/Mass Spectrometry

MRM extracts were analyzed using an Agilent gas chromatograph (Model 6890) equipped with a mass spectrometer detector (MSD, Model 5973N) with an inert source and auto sampler. Standards included: 3-hydroxy-PCB-30 (98.1% purity, purchased from AccuStandard, Inc., New Haven, CT), 4-hydroxy-PCB-30 (100 µg/ml in isooctane, Accustandard), and PCB-30 (100 µg/ml in hexane, Ultrascientific, Kingstown, RI). A ZB-5MS Guardian column (30 m x 0.25 mm i.d., film thickness 0.25 µm, Phenomenex, Torrance, CA) was used for sample separation with Helium as carrier gas at a constant flow rate (1.0 mL/min) maintained by an electronic pressure control module. Mass spectrometric measurements with electron ionization (EI) at 70 eV were performed in selected ion monitoring (SIM) mode. The molecular and fragment ions used for SIM recording were 272 (base peak), 202, 173, and 139. The temperature for MS source and MS Quad were set at 250 and 200 ˚C, respectively. The temperature program started at 190 ˚C and increased to 200 ˚C at 10 ˚C/min, then increased at a rate of 2.5

˚C/min up to 230 ˚C. The injection temperature was set at 250 ˚C and the transfer line was at 280 ˚C.

Results

In Silico SMARTCyp Predictions

PCB-30 is predicted by SMARTCyp to have several reactive sites that can be oxidized by CYP2D6 and CYP3A4 (Table 2.1.1). Carbon atom C.4 (see atom numbering on Figure 2.1.1) is predicted to be the most reactive (i.e., lower energy values in Table 2.1.1) and most accessible in both P450 cases. In the CYP2D6 case, the atoms C.2 and C.3' are predicted to be more reactive than the C.3 atom. However, the C.2 and C.3' atoms are also predicted to be less accessible than the C.3, resulting in a lower overall predicted rank (i.e., a lower score in Table 2.1.1) for C.2 and C.3' than for C.3. In the CYP3A4 case, atom C.3 is predicted to have a lower rank than it does in the CYP2D6 case because of a less favorable accessibility. These results suggest that C.4 atom on PCB-30 is the most likely site of hydroxylation by both CYP2D6 and CYP3A4. However, SMARTCyp predicted the next possible oxidation site of PCB differently for the two P450s: atom C.3 in the case of CYP2D6, and atom C.2 in the case of CYP3A4.

In Silico P450 Docking Predictions

Molecular docking was used as a better method than SMARTCyp to determine accessibility of ligand atoms to the reactive oxygen atom which is bonded to the heme-iron atom. Flexible ligand and protein docking simulations were carried out and the distance of ligand atoms to the reactive oxygen atom of specific P450 enzymes are reported in Table 2.1.2. Binding mode depictions for PCB-30 docked in each P450 can be viewed in Supplemental Figures 2.1.S2-2.1.S6. During initial placement of ligand atoms, 660 unique ligand poses were returned for CYP2D6 and 1021 unique poses were returned for CYP3A4. The C.4 and C.3 atoms for PCB-30 are both equally positioned near the reactive oxygen atom (3.0 Å and 3.0 Å, respectively) in CYP2D6 and (3.4 Å and 3.8 Å, respectively) in CYP3A4. The other two carbon atoms, C.2 and C.3', were found to be ~4-5 Å from the reactive oxygen for both P450s, in contrast with the non-structure-based SMARTCyp results. These docking results suggest that the 3- and 4-hydroxy species of PCB-30 are the most likely metabolites for these two specific P450s investigated here.

In Silico Estrogen Receptor Docking Predictions

Docking of the predicted 4-, 3-, and 3'-hydroxy metabolites for PCB-30 was carried out with the crystal structure of hER-α in the agonist state (1L2I). PCB-30 has previously been shown as only weakly active toward the hER-α protein but increasing in activity after being metabolized into hydroxylated forms[22,23]. The hER-α docking results predict that all of the hydroxylated metabolites of PCB-30 bind more strongly (i.e., lowest predicted binding free energy) to the hormone receptor than the non-hydroxylated PCB-30 parent compound (Table 2.1.3). The 4-hydroxy metabolite exhibits the most favorable binding energy (-11.0 kcal/mol) followed by the 3-hydroxy and 3'-hydroxy metabolites (-8.0 kcal/mol and -8.2 kcal/mol, respectively). It is interesting to

note that the 3'-hydroxy species is predicted to bind slightly better to the estrogen receptor than the 3-hydroxy species (-8.2 kcal/mol vs. -8.0 kcal/mol). In principle, this would suggest that the 3'-hydroxy metabolite is more estrogenic, or at least as estrogenic, than the 3-hydroxy species. However the docking results given in Table 2.1.2 indicate that the 3'-hydroxy metabolite is unlikely to be formed by P450 3A4 and 2D6 metabolism.

Structural examination of the binding pocket interactions (Figure 2.1.2) showed that docking reproduced important protein-ligand interactions at GLU 353 and the adjacent water molecule (water 10).[61] The ortho- and para-hydroxyl groups of 3-hydroxy-PCB-30 and 4-hydroxy-PCB-30 are found to be involved in a H-bond network at these locations. Establishing these H-bonds is thought to contribute to the commonly observed increase in estrogenic activity for hydroxylated molecules.[34,35,61] Additional conserved interactions are H-bonds at ARG 394 and HIS 524 and arene-H bonds between PHE 404, LEU 387, LEU 384 (accepting electrons) and the nearest ligand phenol (donating electrons)[61,62] as shown in Figure 2.1.2.

Table 2.1.1. SMARTCyp: Reactive Atom Sites.

A. CYP2D6

Rank$^{\alpha}$	Atom$^{\theta}$	Score$^{\beta}$	Energy$^{\gamma}$	S2End$^{\varepsilon}$
1	C.4	80.8	80.8	0
2	C.3	93	86.3	1
3	C.2	94.2	80.8	2
4	C.3'	97.5	84.1	2

B. CYP3A4

Rank$^{\alpha}$	Atom$^{\theta}$	Score$^{\beta}$	Energy$^{\gamma}$	Accessibility$^{\delta}$
1	C.4	72.8	80.8	1
2	C.2	74.8	80.8	0.75
3	C.3'	78.1	84.1	0.75
4	C.3	79.3	86.3	0.88

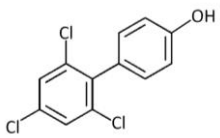
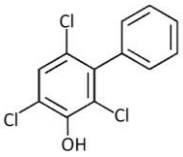
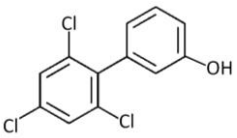
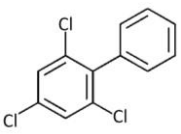
SMARTCyp reactivity (energy) and accessibility (S2End or Accessibility) of carbon atoms in PCB-30 for CYP2D6 (Panel A) and CYP3A4 (Panel B). Carbon atom$^{\theta}$ labels refer to the atom types provided at the top of Figure 2.1.1. The rank$^{\alpha}$ is based on the score$^{\beta}$ which is calculated with a formula[28] combining both reactivity and accessibility. Energy$^{\gamma}$ is determined by calculating the transition states for each atom with the lowest value being the most favorable. In the CYP3A4 case, atom accessibility$^{\delta}$ is determined with highest "Accessibility" values (up to a maximum value of 1) indicating better accessibilities. In the CYP2D6 case, lowest "S2End" values indicate better atomic accessibilities.

Table 2.1.2. P450 Docking (2D6 and 3A4): Atom Accessibility.

Carbon Atom	CYP2D6 Distance (angstroms)	CYP3A4 Distance (angstroms)
C.4	3.0	3.4
C.3	3.0	3.8
C.2	4.0	5.2
C.3'	4.0	4.7

Distance between the CYP450 reactive oxygen atom and the carbon atoms of PCB-30 after docking. Carbon atom numbers refer to the labels provided at the top of Figure 2.1.1

Table 2.1.3. Docking (hER-α): Binding Predictions for PCB-30 and Metabolites.

Ligands	**Structure**	**London dG Score**(kcal/mol)
4-hydroxy-PCB-30		-11.0
3'-hydroxy-PCB-30		-8.2
3-hydroxy-PCB-30		-8.0
PCB-30		-7.2

Ligand Names, structures and London dG binding scores for each compound docked in hER-α.

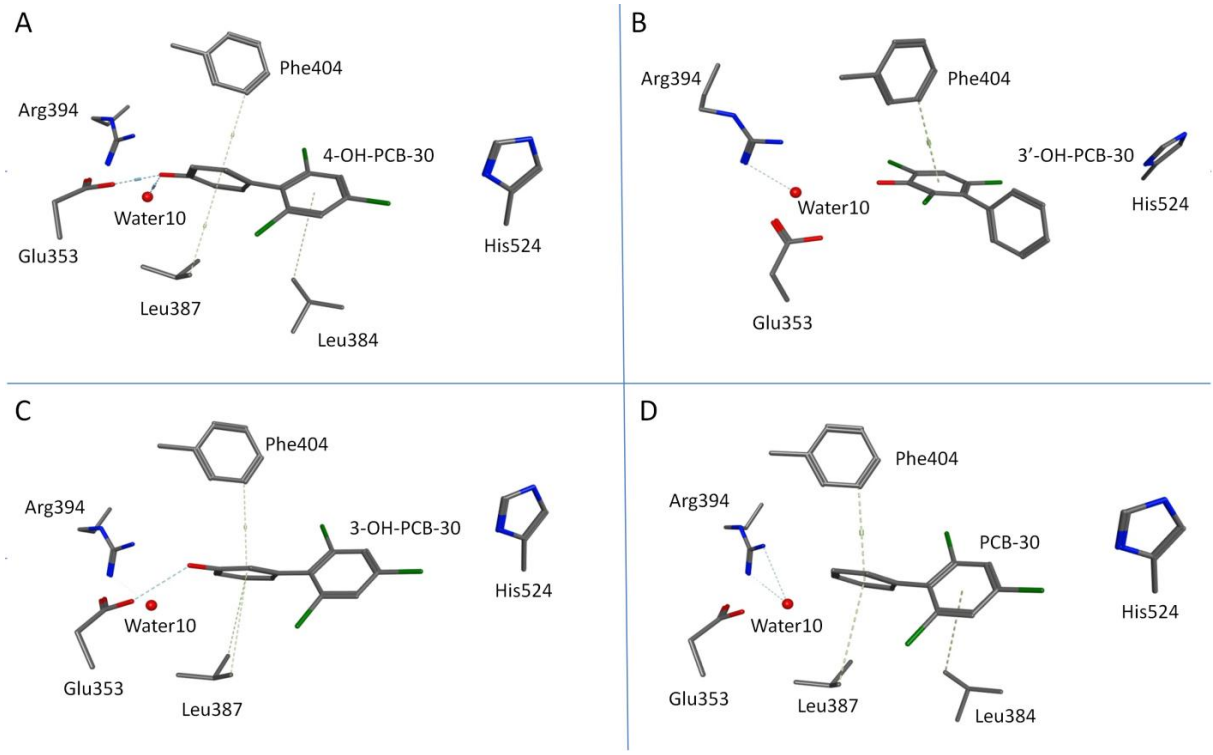


Figure 2.1.2. Docking (hER-α): Atom Interactions. Protein-ligand interactions for PCB-30 and its metabolites docked in hER-α. Only interacting residues are shown. Hydrogen and backbone atoms are hidden. The dotted lines represent noncovalent interactions. A: (4-hydroxy-PCB-30). B: (3'-hydroxy-PCB-30). C: (3-hydroxy-PCB-30). D: (PCB-30).

In Vitro P450 Exposure and Bioassays

S. cerevisiae BLYES and BLYR were incubated with PCB-30 and 4-hydroxy-PCB-30, yielding EC_{50} values of 4.2×10^{-5} M and 3.4×10^{-8} M, indicating that addition of the hydroxyl group at position C.4 on PCB-30 resulted in an approximately 1000x increase in estrogenic response (Figure 2.1.3). To explore whether cytochrome P450 metabolism also resulted in increased activity, PCB-30 was incubated with total liver microsomes, enriched CYP2D6 extract, and enriched CYP3A4 extract. Incubation for 24 hours with any of the microsomal or enriched P450 extracts resulted in an 800,000-1,000,000 fold increase in hER-α activity, indicating that liver P450s were capable of hydroxylating PCB-30 into a more estrogenic form (Figure 2.1.4).

Gas Chromatography/Mass Spectrometry

GC/MS was used in order to validate that 3- and 4-hydroxy metabolites were produced in each of the respective MRMs (Figure 2.1.5). The retention times for 3-hydroxy-PCB-30 and 4-hydroxy-PCB-30 standards were 9.083 and 9.406 min, respectively. Peaks at both of these retention times were detected in the 3A4 and 2D6 MRMs. The ratio based on peak areas between 3-hydroxy- and 4-hydroxy-PCB-30 was approximately 1:19 with the 4-hydroxy metabolite being more abundant than the 3-hydroxy metabolite.

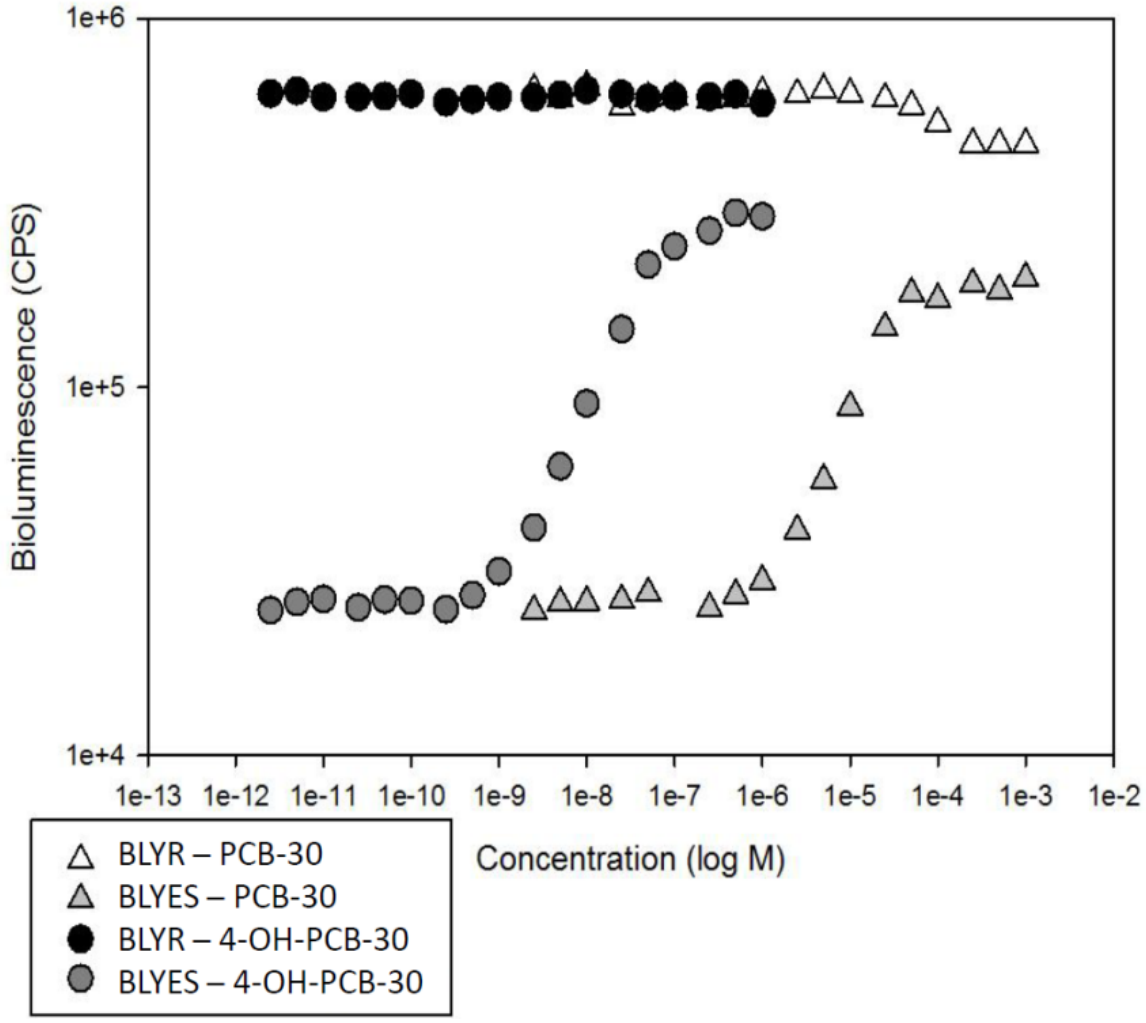


Figure 2.1.3. Bioassay: Response to Standards (PCB-30 and 4-hydroxy-PCB-30). Dose-response curves for standards (PCB-30 and 4-OH-PCB-30). Estrogenic response (bioluminescence) was measured in the BLYES and BLYR assays.

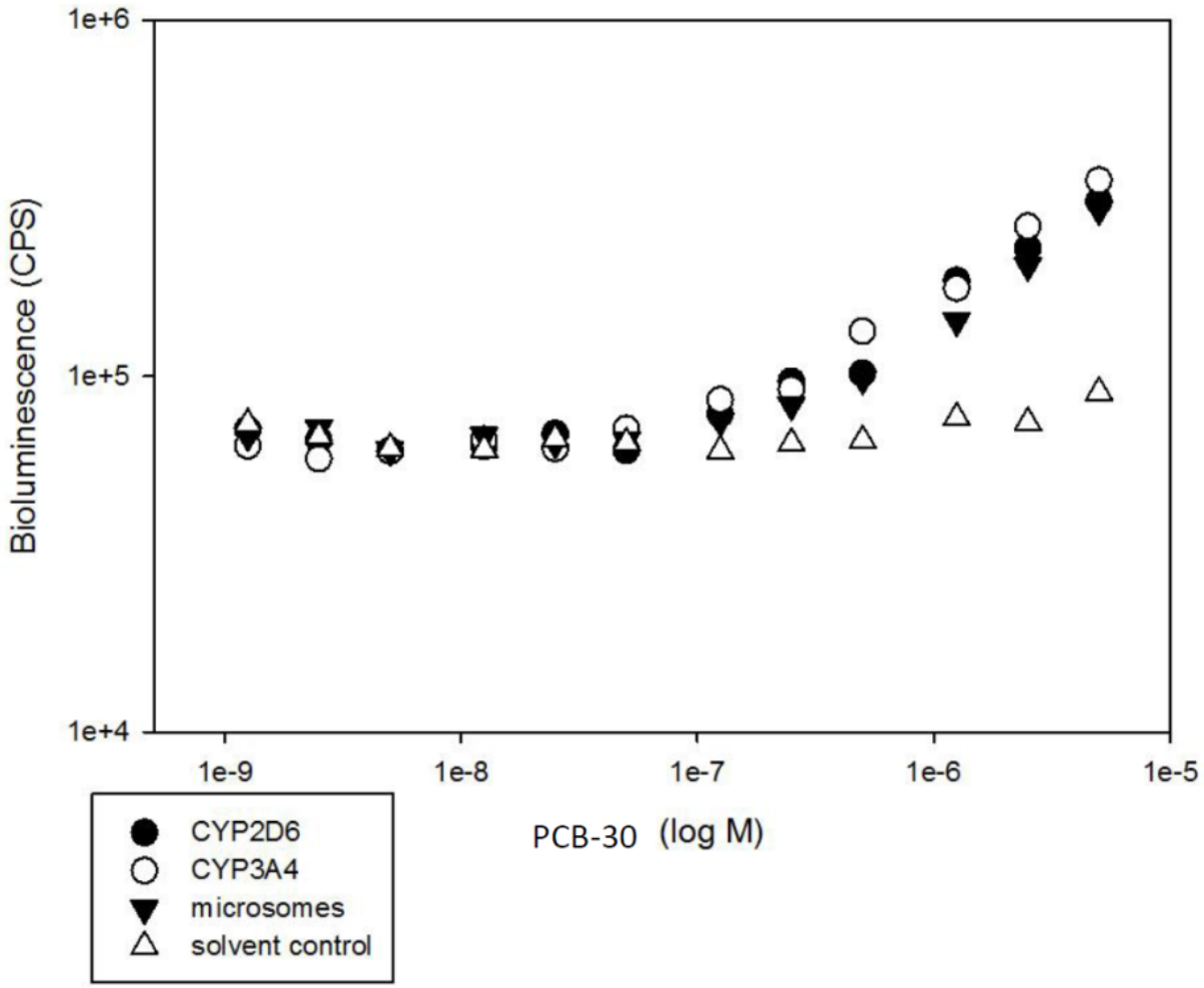


Figure 2.1.4. Bioassay: Response to PCB-30 Metabolites (MRMs). Response to PCB-30 Metabolites (MRMs). Shown is the estrogenic response (bioluminescence) of PCB-30 MRMs that were generated in either CYP2D6, CYP3A4, total liver microsome, or a solvent control.

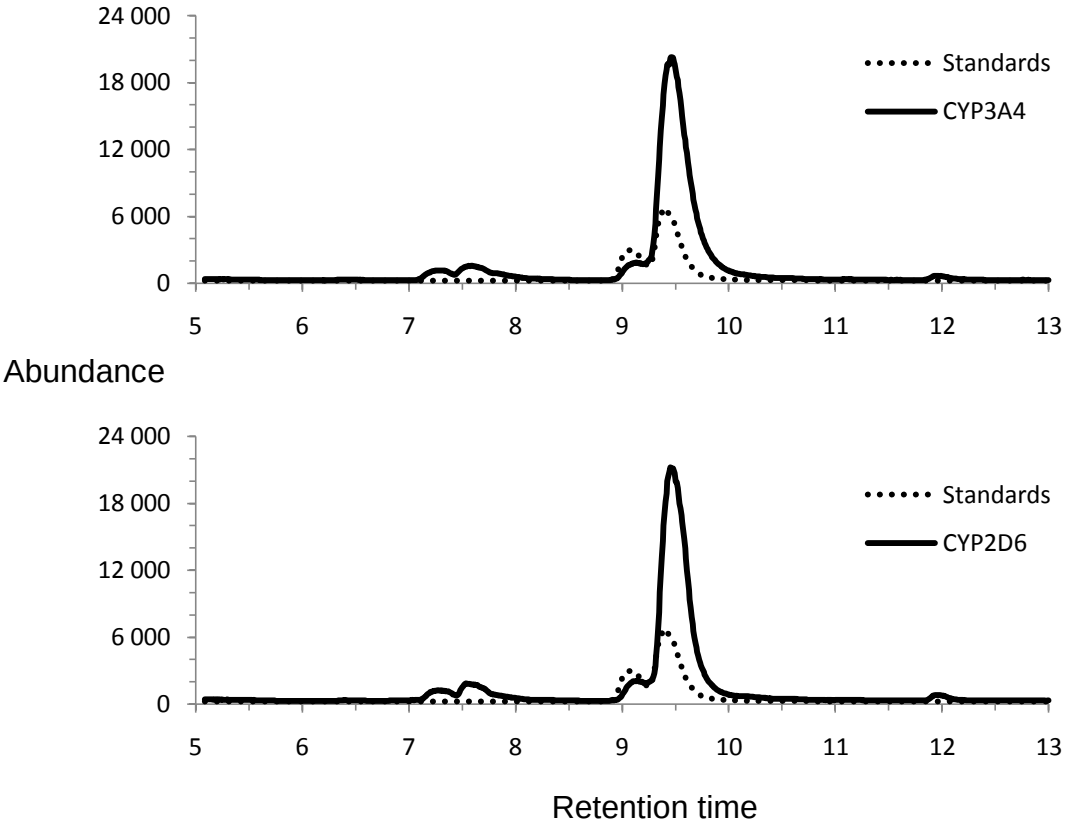


Figure 2.1.5. GC/MS: Characterization of PCB-30 Metabolites (MRMs). GC/MS Characterization of PCB-30 Metabolites (MRMs). Dotted black line: peak retention times for the two standards (3-hydroxy-PCB-30 and 4-hydroxy-PCB-30). The 3-hydroxy peak is approximately 9.1 min and the 4-hydroxy peak is approximately 9.4 min. Solid black lines: peak retention times for PCB-30 metabolites formed during incubation with CYP3A4 (Top Panel) and CYP2D6 (Lower Panel).

Discussion

The reactivity scoring obtained by SMARTCyp calculations predicts that the 4-, 3-, 3'-, and 2-carbon atoms are reactive and could potentially become hydroxylated. Inclusion of SMARTCyp's 2-dimensional accessibility metrics predicts the top two sites of metabolism as C.4 and C.3 for CYP2D6 and C.4 and C.2 for CYP3A4. However, docking predicts the C.4 and C.3 atoms as the most accessible and likely oxidation sites for both P450s. GC/MS results have only two peaks with matching retention times to the 4- and 3-hydroxy-PCB-30 standards which confirms that the docking predictions are more accurate than SMARTCyp when factoring in atom accessibility. The bimodal GC/MS spectrum suggests that there were no more than two metabolites, unless 3- and 3'-hydroxy-PCB-30 share the same retention time. However, there is no mention of 3'-

hydroxy-PCB-30 as a known EDC in the literature which suggests that it may only be a theoretical derivative of PCB-30 and unlikely to naturally occur.

It is interesting to note that the predictive methods from this model were able to aid in the strategic selection of reference spectra for validating the metabolites being produced by these two specific P450s. The model's predictions suggested that PCB-30 metabolites in CYP2D6 and CYP3A4 were likely the 3- and 4-hydroxy species. We used this information to rationalize using 3- and 4-hydroxy-PCB-30 compounds as standards for producing reference spectra during the GC/MS analysis. Otherwise, selection of standards for GC/MS would have been a challenge requiring the referencing of many more spectra. This was not an intended application for the model but it is an interesting benefit of this technology.

Hydroxylation of PCB-30 has previously been shown to lead to an increased estrogenic response and that 4-hydroxy-PCBs are a major P450 metabolite class of PCBs.[15,22,23] The present study expands upon these previous findings by identifying CYP2D6 and CYP3A4 as specific P450s involved in the ortho- or para- hydroxylation of PCB-30 into its more estrogenic forms. The present results do not exclude the involvement of other CYP450s in estrogenizing PCB-30 and only test estrogenic activity with hER-α in an agonist state. The estrogen receptor also exists in the antagonist state and there are at least two other estrogen receptor isotypes (beta and gamma).[65] These additional forms of hER-α would eventually be important to consider for more complex screening of EDC activity. Also to consider are other cytochrome P450 isoforms that may metabolize EDCs.[30,33,12] The reliance of this model on virtual docking means that only protein structures are required to include these additional proteins, which many already exist or can be homology modeled. Also, SMARTCyp has recently been expanded to include reactivity predictions for several additional CYP450s (1A2, 2A6, 2B6, 2C8, 2C19 and 2E1).

Conclusion

This work established that CYP3A4 and CYP2D6 produce PCB-30 metabolites with higher estrogenic activity than their parent structure. Two specific metabolites were produced that had been modified to contain a hydroxyl group at either the C.4 or C.3 atoms with 4-hydroxy-PCB-30 being the primary product and the most estrogenic of observed metabolites. All of these results were able to be predicted using a combined reactivity-based (SMARTCyp) and multi-structure-based approach (multi-protein docking). This integrated model uses techniques which were individually developed to provide less costly and more efficient compound characterizations toward specific proteins, and integration of these techniques allows for robustly predicting compounds involved in both P450 metabolism and estrogen activity. Based on the scalability and adaptability of this combined approach, it is conceivable that this method can be further developed and used to robustly predict the P450 metabolism and estrogenic effects of many more environmental and pharmaceutically important compounds. This work may also serve as an example for future attempts at modeling the metabolism and bioactivity of other compound and protein classes through increasingly complex multi-protein pathways.

Acknowledgment

This work was financially supported by a start-up grant from the University of Tennessee to J.B. J.H acknowledges support by the Genome Science and Technology graduate school and the IGERT: SCALE-IT fellowship (NSF Award 0801540).

Abbreviations

CYP450, Cytochrome P450; P450, Cytochrome P450; CYP2D6, Cytochrome P450 2D6; CYP3A4, Cytochrome P450 3A4; PCB, Polychlorinated Biphenyl; MRM, Microsomal Reaction Mixture; hER-α, Human Estrogen Receptor Alpha; EDC, Endocrine Disrupting Chemical; QSAR, Quantitative Structure Activity Relationship; BLYES, Bioluminescent Yeast Estrogen Strain; PAHs, Polynuclear Aromatic Hydrocarbons; MOE, Molecular Operating Environment; RMS, Root Mean Square; GB/VI, Generalized *Born*/Volume Integral, Generalized Born; PDB, Protein Database; GLU, Glutamate; ARG, Arginine; HIS, Histidine; PHE, Phenylalanine; LEU, Leucine; E_place, Placement Score; BLYR, Constitutive Bioluminescent Yeast Strain Gas Chromatography/Mass Spectrometry; GC/MS.

Supplemental Figures

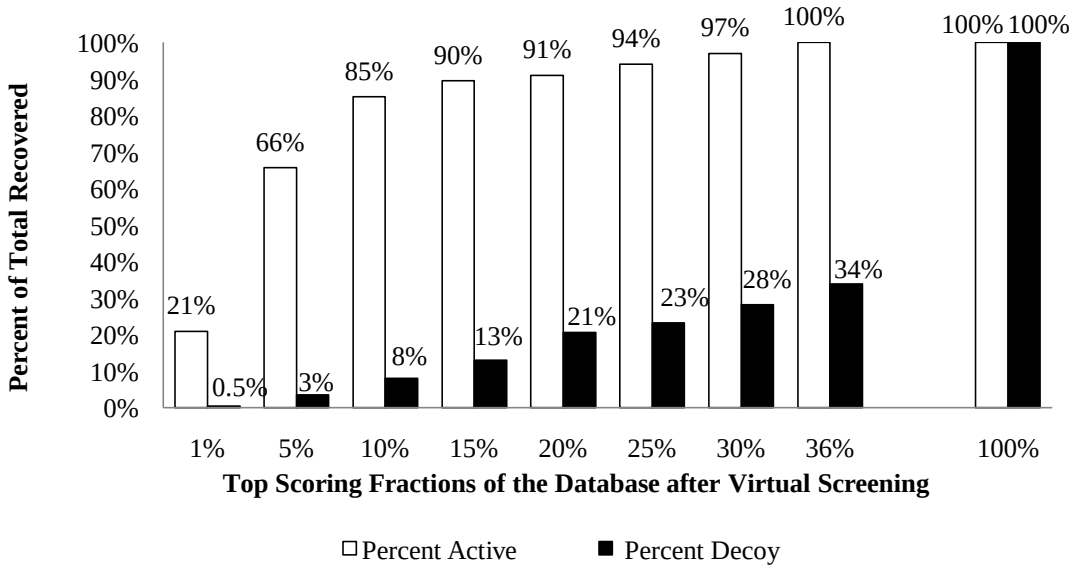


Figure 2.1.S1. Computational Prioritization of the DUD-ER Database. This graph displays the docking results for the Directory of Useful Decoy (DUD) database. There are 67 known active molecules and 2569 known decoys in this database.

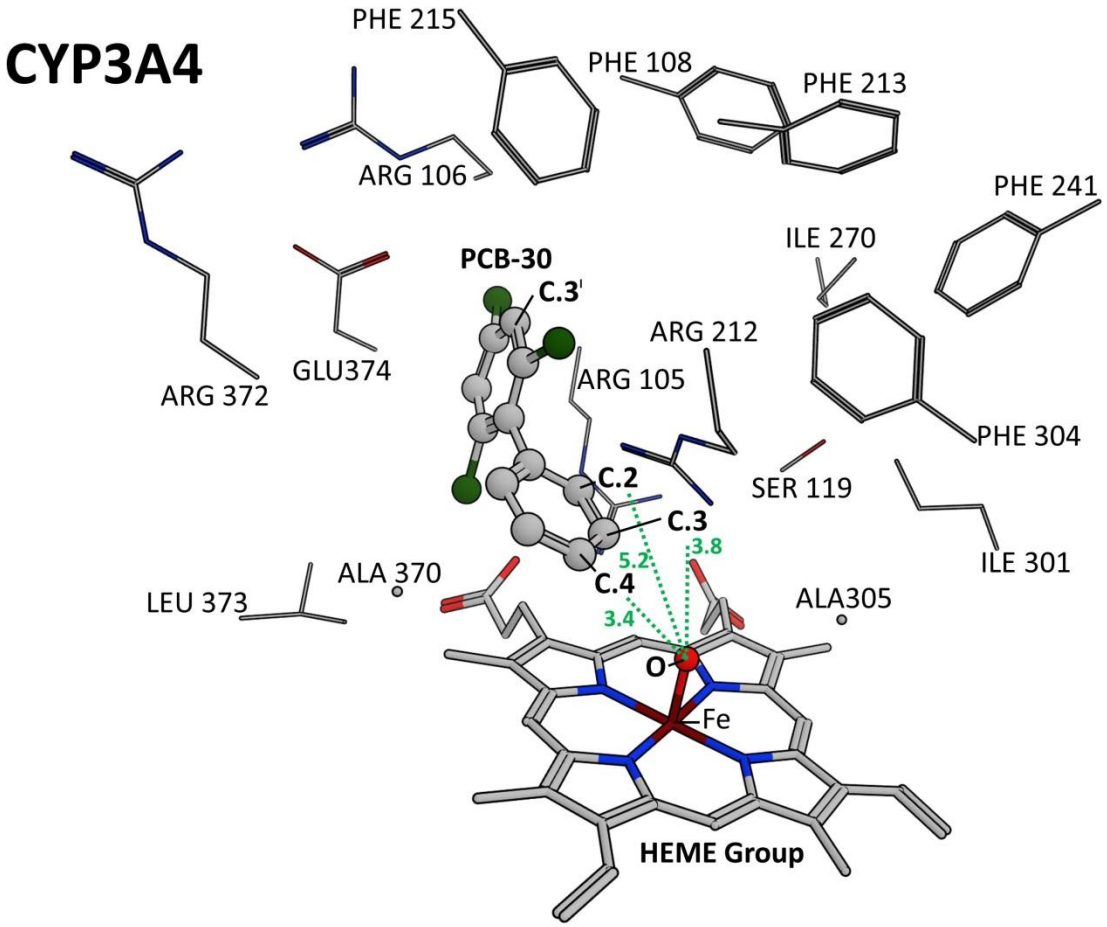


Figure 2.S2. Docking Pose for PCB-30 Atoms C.4, C.3, C.2 in CYP3A4. Shown is the docking pose of PCB-30 with atoms C.4, C.3, and C.2 closest (most accessible) to the reactive oxygen atom on the CYP3A4 heme group. Dotted green lines and green text show distance in angstroms from respective PCB-30 atoms to the heme oxygen atom. Sidechain atoms within 4.5 Å of PCB-30 are made visible and labeled to depict the binding pocket.

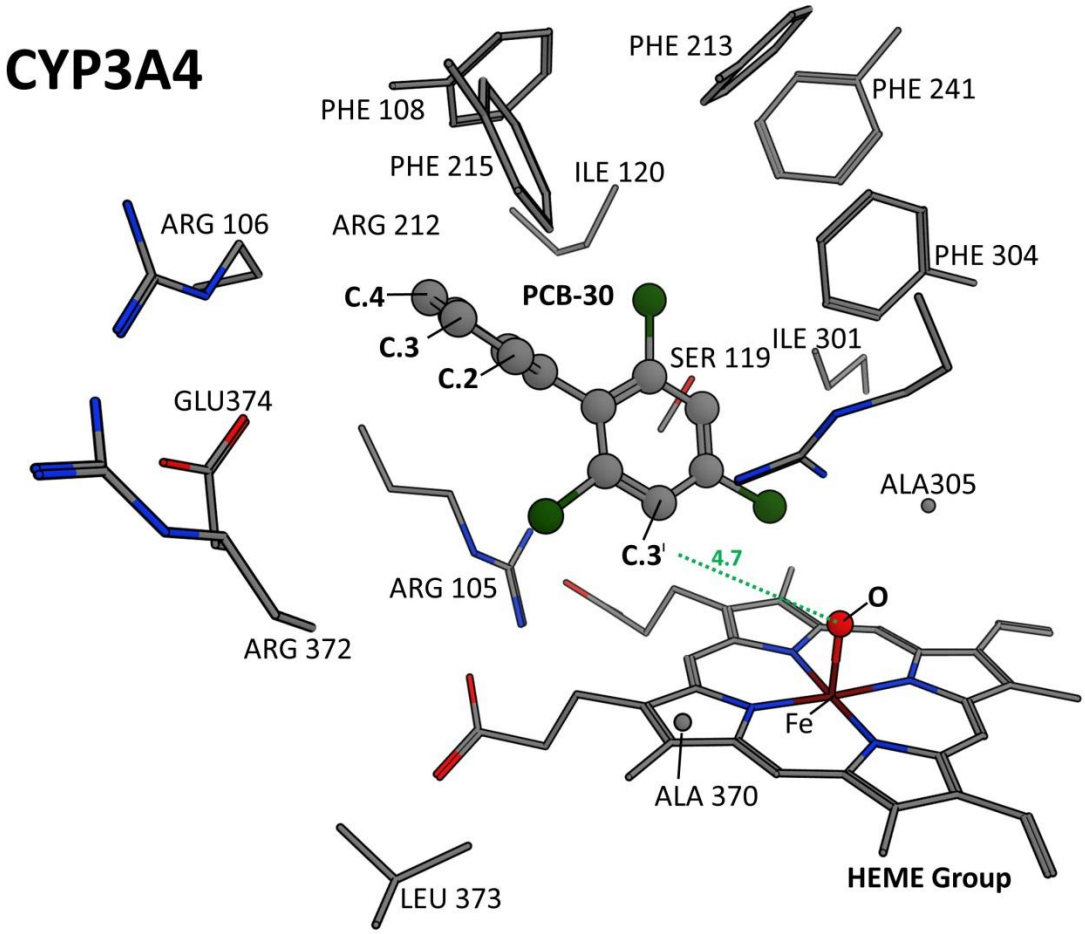


Figure 2.1.S3. Docking Pose for PCB-30 Atoms C.3' in CYP3A4. Shown is the docking pose of PCB-30 with atom C.3' closest (most accessible) to the reactive oxygen atom on the CYP3A4 heme group. Dotted green lines and green text show distance in angstroms from respective PCB-30 atoms to the heme oxygen atom. Sidechain atoms within 4.5 Å of PCB-30 are made visible and labeled to depict the binding pocket.

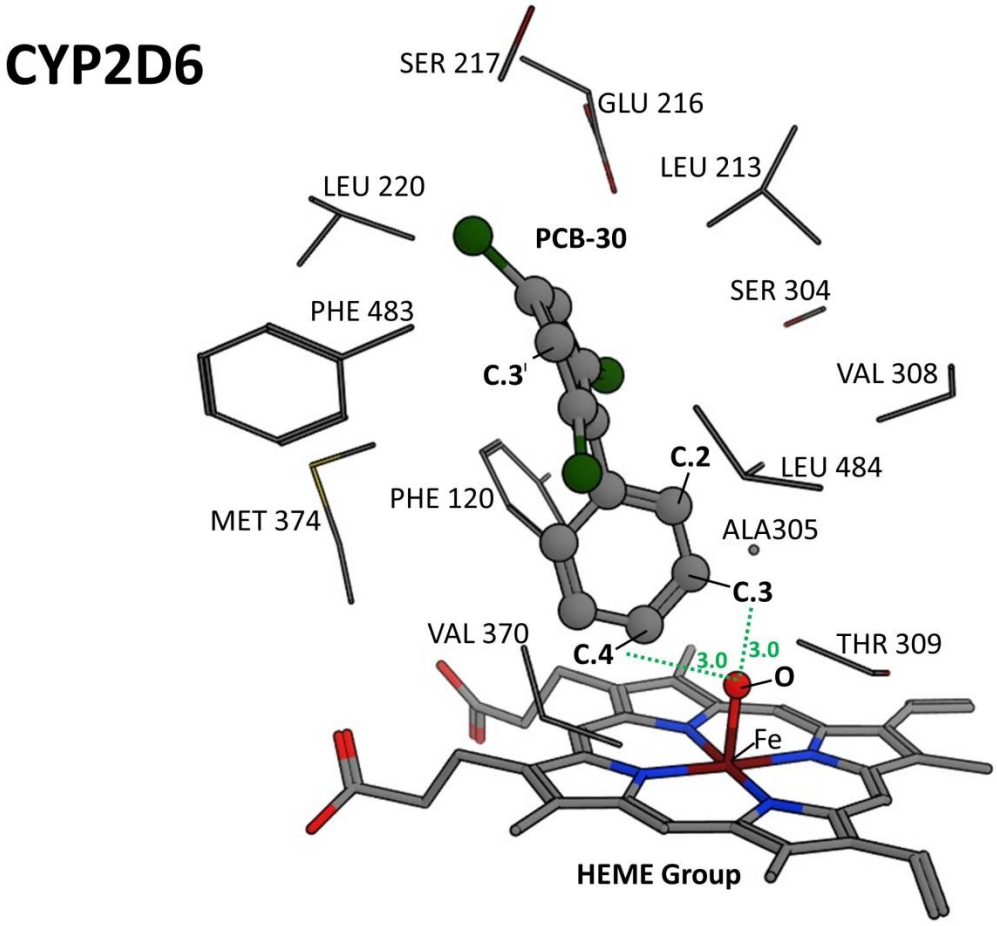


Figure 2.1.S4. Docking Pose for PCB-30 Atoms C.4 and C.3 in CYP2D6. Shown is the docking pose of PCB-30 with atoms C.4 and C.3 closest (most accessible) to the reactive oxygen atom on the CYP2D6 heme group. Dotted green lines and green text show distance in angstroms from respective PCB-30 atoms to the heme oxygen atom. Sidechain atoms within 4.5 Å of PCB-30 are made visible and labeled to depict the binding pocket.

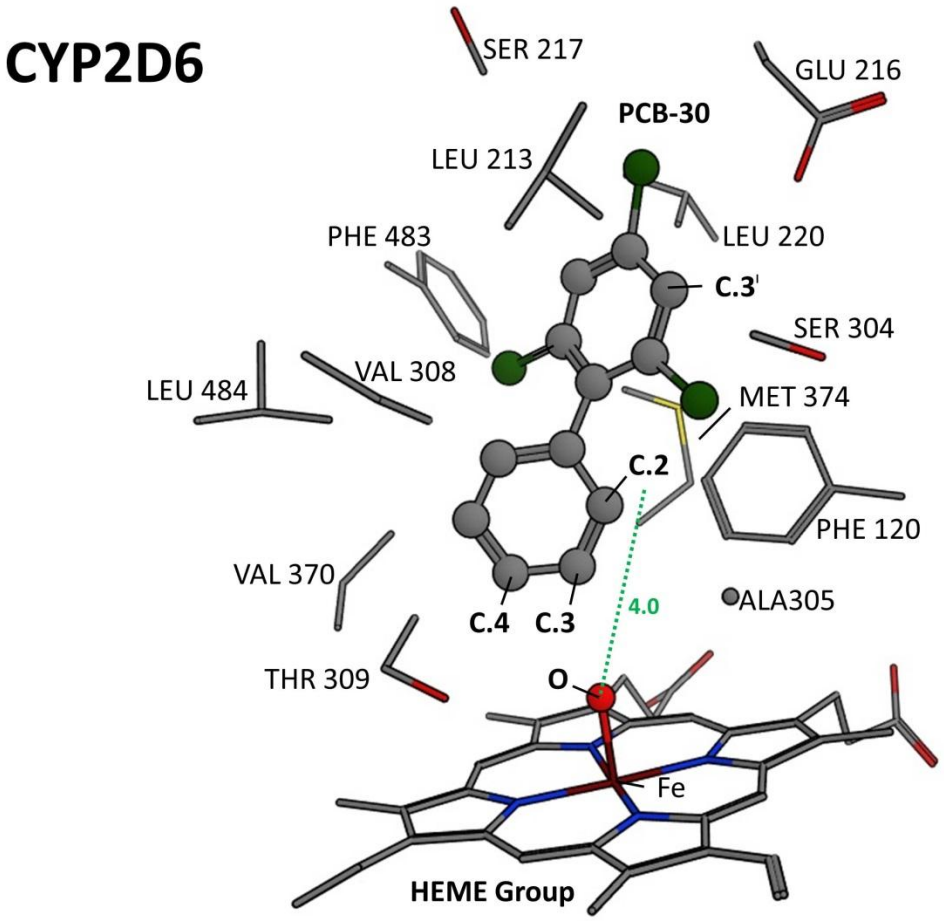


Figure 2.1.S5. Docking Pose for PCB-30 Atom C.2 in CYP2D6. Shown is the docking pose of PCB-30 with atom C.2 closest (most accessible) to the reactive oxygen atom on the CYP2D6 heme group. Dotted green lines and green text show distance in angstroms from respective PCB-30 atoms to the heme oxygen atom. Sidechain atoms within 4.5 Å of PCB-30 are made visible and labeled to depict the binding pocket.

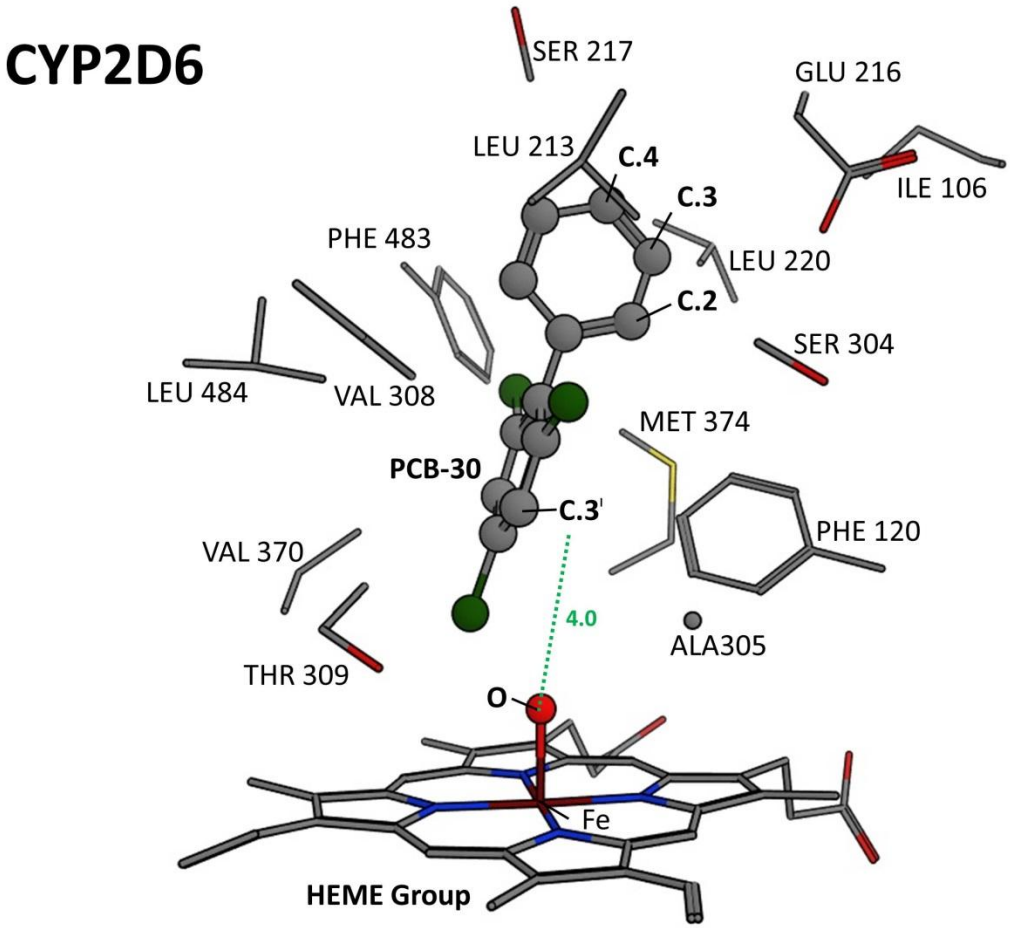


Figure 2.1.S6. Docking Pose for PCB-30 Atom C.3' in CYP2D6. Shown is the docking pose of PCB-30 with atom C.3' closest (most accessible) to the reactive oxygen atom on the CYP2D6 heme group. Dotted green lines and green text show distance in angstroms from respective PCB-30 atoms to the heme oxygen atom. Sidechain atoms within 4.5 Å of PCB-30 are made visible and labeled to depict the binding pocket.

References

(1) National Institute of Environmental Health Sciences. 2010. Endocrine disruptors. [cited 2012 March 21]. Available from: http://www.niehs.nih.gov/health/assets/docs_a_e/endocrine_disruptors.pdf.

(2) Diamanti-Kandarakis E, Bourguignon J-P, Giudice LC, Hauser R, Prins GS, Soto AM, Zoeller TR, Gore AC. 2009. Endocrine-disrupting chemicals: an endocrine society scientific statement. *Endocrine Rev*. 30:293–342.

(3) US Congress. 1996. Food Quality Protection Act, Pub. L. No. 104-170 (August 3, 1996). Washington, DC.

(4) US Congress. 1996. Safe Drinking Water Act Amendments of 1996, Pub. L. 104-182 (August 6, 1996). Washington, DC.

(5) US Environmental Protection Agency. 1998. Endocrine Disruptor Screening and Testing Advisory Committee final report. EPA/743/R-98/003. Washington, DC.

(6) Svendsgaard, D.J., Ward, T.R., Tilson, H.A., Kodavanti, P.R.S. 1997. Empirical modeling of an in vitro activity of polychlorinated biphenyl congeners and mixtures. *Environmental health perspectives*. 105, 1106-1115.

(7) Arulmozhiraja, S., Shiraishi, F., Okumura, T., Iida, M., Takigami, H., Edmonds, J.S., Morita, M. 2005. Structural Requirements for the interaction of 91 hydroxylated polychlorinated biphenyls with estrogen and thyroid hormone receptors. *Toxicol. Sci*. 84, 49-62.

(8) Layton, A.C., Sanseverino, J., Gregory, B.W., Easter, J.P., Sayler, G.S., Schultz, W.T. 2002. In vitro estrogen receptor binding of PCBs: measured activity and detection of hydroxylated metabolites in a recombinant yeast assay. *Toxicology and Applied Pharmacology*. 180, 157-63.

(9) Li ,Y., Harner, T., Liu, L., Zhang, Z., Ren, N., Jia, H., Ma, J., Sverko, E.. 2010. Polychlorinated biphenyls in global air-soil exchange, and fractionation effect. *Environ. Sci. Technol.*44, 2784-2790.

(10) UNEP. 2001. Regionally based Assessment of Persistent Toxic Substances: Central and North East Asia Regions; United Nation Environmental Program: Nairobi, Kenya, No. 10.

(11) Fernandez, M.F., Kiviranta, H., Molina-Molina, J.M., Lain, O., Lopez-Espinosa, M.J., Vartiainen, T., Olea, N. 2008. Polychlorinated biphenyls (PCBs) and hydroxyl-PCBs in adipose tissue of women in southeast Spain. *Chemosphere*. 71, 1196-1205.

(12) Berg, V., Lyche, J.L., Gutleb, A.C., Lie, E., Skaare, J.U., Aleksandersen, M., Ropstad, E. 2010. Distribution of PCB 118 and PCB 153 and hydroxylated PCB metabolites (OH-CBs) in maternal, fetal and lamb tissues of sheep exposed during gestation and lactation. *Chemosphere*. 80, 1144-1150.

(13) Beyer, A., and Biziuk, M. 2009. Environmental fate and global distribution of polychlorinated biphenyls. *Reviews of Environ. Contamination and Toxicology*. 201, 137-158.

(14) Arulmozhiraja, S., Shiraishi, F., Okumura, T., Iida, M., Takigami, H., Edmonds, J.S., Morita, M. 2004. Structural requirements for the interaction of 91 hydroxylated polychlorinated biphenyls with estrogen and thyroid hormone receptors. *Toxicological Sciences.* 84, 49-62.

(15) Vakharia, D., Gierthy, J. 1999. Rapid assay for oestrogen receptor binding to PCB metabolites. *Toxicology In Vitro*. 13, 275–282.

(16) Gozgit, J.M., Nestor, K.M., Fasco, M.J., Pentecost BT, Arcaro KF. 2004. Differential action of polycyclic aromatic hydrocarbons on endogenous estrogen-responsive genes and on a transfected estrogen -responsive reporter in MCF-7 cells.*Toxicology and Applied Pharmacology* 196, 58-67.

(17) Delaforge, M., Pruvost, A., Perrin, L., Andre, F. 2005. Cytochrome P450-mediated oxidation of glucuronide derivatives: example of estradiol-17β-glucuronide oxidation to 2-hydroxy-estridiol-17β-glucuronide by CYP 2C8. *Drug Metabolism & Disposition* 33, 466-473.

(18) Carson, D.B., and Williams, D.E.. 2001. 4-hydroxy-2',4',6'-trichlorobiphenyl and 4-hydroxy-2',3',4',5'-tetrachlorobiphenyl are estrogenic in rainbow trout. *Environ. Toxicol. Chem.* 20, 351-358.

(19) Desta, W., Soukhova, F. 2004. Comprehensive evaluation of tamoxifen sequential biotransformation by the human cytochrome P450 system in vitro: prominent roles for CYP3A and CYP2D6. *J. of Pharm. and Experimental Therap.* 310, 1062-1075.

(20) McGraw, J.E., and Waller, D.P. 2009. The role of African American ethnicity and metabolism in sentinel polychlorinated biphenyl congener serum levels. *Environmental Toxicology and Pharmacology* 27, 54–61.

(21) Huang, P., Zhu, B., Wang, L.S., Ouyang, D.S., Huang, S.L., Chen, X.P., Zhou, H.H. 2003. Relationship between CYP3A activity and breast cancer suspeptibility in chinese women. *Eur. J. Clinical Pharmacol*. 59, 471-476.

(22) Sandala, G.M., Sonne-Hansen, C., Dietz, R., Muir, D.C.G., Valters, K., Bennett, E.R., Born, E.W., Letcher, R.J. 2004. Hydroxylated and methyl sulfone PCB metabolites in adipose and whole blood of polar (*Ursus maritimus*) from east Greenland. *Science of the Total Environment* 331, 125-141.

(23) Gilroy, E.A.M., Muir, D.G.C., McMaster, M.E., Darling, C., Campbell, L.M., De Solla, S.R., Parrott, J.L., Brown, S.B., Sherry, J.P. 2012. Polychlorinated biphenyls and their hydroxylated metabolites in wild fish from Wheatley Harbor area of concern, Ontario, Canada. *Environmental Toxicology and Chemistry* 31, 2788-2797.

(24) Li, .F, Xie, Q., Li, X., Li, N., Chi, P., Chen, J., Wang, Z., Hao, C. 2010. Hormone activity of hydroxylated polybrominated diphenyl ethers on human thyroid receptor-

beta: in vitro and in silico investigations. *Environmental Health Perspectives 118*, 602–606.

(25) Rydberg, P., Gloriam, D.E., Zaretzki, J., Breneman, C., Olsen, L. 2010. SMARTCyp: A 2D method for prediction of cytochrome P450-mediated drug metabolism. *ACS Med. Chem.Lett.* 1, 96-100.

(26) Schapira, M., Abagyan, R., Totrov, M. 2003. Nuclear Hormone Receptor Targeted Virtual Screening. *J. Med. Chem*. 46, 3045-3059.

(27) Nose, T., Tokunaga, T., Shimohigashi, Y. 2009. Exploration of endocrine-disrupting chemicals on estrogen receptor alpha by the agonist/antagonist differential-docking screening (AADS) method: 4-(1-Adamantyl)phenol as a potent endocrine disruptor candidate. *Toxicology Letters* 191, 33-39.

(28) Hritz, J., de Ruiter, A., Oostenbrink, C. 2008. 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, 7469-7477.

(29) Baudry, J., Rupasinghe, S., Schuler, M. 2006. Class-dependent sequence alignment strategy improves the structural and functional modeling of P450s. *Protein Eng. Des. Sel.* 19, 345-353.

(30) Baudry, J., Li, W., Pan, L., Berenbaum, M.R., Schuler, M.A. 2003. Molecular docking of substrates and inhibitors in the catalytic site of CYP6B1, and insect cytochrome P450 monooxygenase. *Protein Engineering* 16, 577-587.

(31) Ginex, T., Cozzini, P., Dall'asta, C. 2014. Preliminary Hazard Evaluation of Androgen Receptor-Mediated Endocrine-Disrupting Effects of Thioxanthone Metabolites through Structure-Based Molecular Docking. *Chemical Research in Toxicology 27*, 279–289.

(32) Ingelman-Sundberg, M. 2005. Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. *Pharmacogenomics J.* 5, 6-13.

(33) Guengerich, F.P. 2003. Cytochromes P450, drugs, and diseases. *Mol Interv.* 3, 194–204.

(34) Kitamura, S., Shimizu, Y., Shiraga, Y., Yoshida, M., Sugihara, K., Ohta, S. 2002. Reductive Metabolism of p, p′-DDT and o, p′-DDT by Rat Liver Cytochrome P450.*Drug metabolism and disposition 30,* 113–118.

(35) Dehal SS, and Kupfer D. 1997. CYP2D6 catalyzes tamoxifen 4-hydroxylation in human liver. *Cancer research* 57:3402-3406.

(36) Vandenbrink, B.M., and Isoherranen, N. 2010. The role of metabolites in predicting drug-drug interactions: Focus on irreversible P450 inhibition. *Current Opinion in Drug Discovery & Development*. 13, 66-77.

(37) Ohkura, K., Kawaguchi, Y., Watanabe, Y., Masubuchi, Y., Shinohara, Y., Hori, H. 2009. Flexible structure of cytochrome P450: promiscuity of ligand binding in the CYP3A4 heme pocket. *Anticancer Research* 29, 935–942.

(38) Goetz, M.P., Rae, J.M., Suman, V.J., Safgren, S.L., Ames, M.M., Visscher, D.W., Reynolds, C., Couch, F.J., Lingle, W.L., Flockhart, D.A., Desta, Z., Perez, E.A., Ingle, J.N. 2005. Pharmacogenetics of tamoxifen biotransformation is associated with clinical outcomes of efficacy and hot flashes. *J. of Clinical Oncology* 23, 9312-9318.

(39) Routledge, E.J., and Sumpter, J.P. 1996. Estrogenic activity of surfactants and some of their degradation products assessed using a recombinant yeast screen. *Environmental Toxicology and Chemistry* 15, 241-248.

(40) Sanseverino, J., Eldridge, M.L., Layton, A.C., Easter, J.P., Yarbrough, J., Schultz, T.W., Sayler, G.S. 2009. Screening of potentially hormonally active chemicals using bioluminescent yeast bioreporters. *Toxicological Sciences* 107, 122-134.

(41) Layton, A.C., Gregory, B.W., Seward, J.R., Schultz, W., Sayler, G.S. 2000. Mineralization of steroidal hormones by biosolids in wastewater treatment systems in Tennessee USA. *Environmental Science & Technology* 34, 3925-3931.

(42) Schultz, T.W., Kraut, D.H., Sayler, G.S., Layton, A.C. 1998. Estrogenicity of selected biphenyls evaluated using a recombinant yeast assay. *Environmental Toxicology and Chemistry* 17, 1727-1729.

(43) Schultz, T.W. 2002. Estrogenicity of biphenylols: activity in the yeast gene activation assay. *Bulletin of Environmental Contamination and Toxicology*. 68, 332-338.

(44) Schultz, T.W., and Sinks, G.D. 2002. Xenoestrogenic gene exression: structural features of active polycyclic aromatic hydrocarbons. *Environmental Toxicology and Chemistry* 21, 783-786.

(45) Sohoni, P., Lefevre, P.A., Ashby, J., Sumpter, J.P. 2001. Possible androgenic/anti-androgenic activity of the insecticide fenitrothion. *Journal of Applied Toxicology* 21, 173-178.

(46) Thomas, K.V., Hurst, M.R., Matthiessen, P., McHugh, M., Smith, A., Waldock, M.J. An assessment of in vitro androgenic activity and the identification of environmental androgens in United Kingdom estuaries. *Environmental Toxicology and Chemistry* 21, 1456-1461.

(47)Raman, D.R., Williams, E.L., Layton, A.C., Burns, R.T., Easter, J.P., Daugherty, A.S., Mullen, M.D., Sayler, G.S. 2004. Estrogen content of dairy and swine wastes. *Environmental Science & Technology* 38, 3567-3573.

(48) Bergamasco, A.M., Eldridge, M., Sanseverino, J., Sodre, F.F., Montagner, C.C., Pescara, I.C., Jardim, W.F., Umbuzeiro, G.A. 2011. Bioluminescent yeast estrogen assay (BYLES) as a sensitive tool to monitor surface and drinking water for estrogenicity. *J. Environ. Monit*. 13, 3288-3293.

(49) Jardim, W.F., Montagner, C.C., Pescara, I.C., Umbuzeiro, G.A., Bergamasco, A.M.B, Eldridge, M.L., Fabriz Sodré, F. 2012. An integrated approach to evaluate emerging contaminants in drinking water. *Separation and Purification Technology* 84, 3-8.

(50) Ellingson, S.R., Smith, J.C., Baudry, J. 2013. VinaMPI: facilitating multiple receptor high-throughput virtual docking on high performance computers. *Computers. J Comp Chem*. 34, 2212–2221.

(51) Borhani, D.W., and Shaw, D.E. 2012. The future of molecular dynamic simulation in drug discovery. *J Comput Aided Mol Des*. 26, 15–26.

(52) *Molecular Operating Environment (MOE)*, 2010.10; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, **2010**.

(53) Labute, P. 2007. Protonate 3D: assignment of macromolecular protonation state and geometry. [cited 1 of March 2013]. Available from: http://www.chemcomp.com/journal/proton.htm

(54) Yang, X., Xie, H., Chen, J., Li, X.. 2013. Anionic phenolic compounds bind stronger with transthyretin than their neutral forms: nonnegligible mechanisms in virtual screening of endocrine disrupting chemicals. *Chemical Research in Toxicology 26*, 1340–7.

(55) Halgren, T.A. 1999. OBForceFieldMMFF94. *J. Comput. Chem.* 20, 720-729.

(56) Yano, J.K., Wester, M.R., Schoch, G.A., Griffin, K.J., Stout, C.D., Johnson, E.F. 2004. The Structure of human microsomal cytochrome P450 3A4 determined by x-ray crystallography to 2.05-Å resolution. *Journal of Biological Chemistry* 279, 38091-38094.

(57) Rowland, P., Blaney, F.E., Smyth, M.G., Jones, J.J., Leydon, V.R., Oxbrow ,A.K., Lewis, C.J., Tennant, M.G., Modi, S., Eggleston, D.S., Chenery, R.J., Bridges, A.M. 2006. Crystal Structure of human cytochrome P450 2D6. *Journal of Biological Chemistry* 281, 7614-7622.

(58) Labute, P. 2008. The generalized Born/volume integral implicit solvent model: estimation of the free energy of hydration using London dispersion instead of atomic surface area. *J. Comput. Chem.* 29, 1693-1698.

(59) Shiau, A.K., Barstad, D., Radek, J.T., Meyers, M.J., Nettles KW, Katzenellenbogen BS, Katzenellenbogen JA, Agard DA, Greene GL. 2002. Structural characterization of a subtype-selective ligand reveals a novel mode of estrogen receptor antagonism. *Nature Structural Biology* 9, 359-364.

(60) Huang, N., Shoichet, B.K., Irwin, J.J. 2006. Benchmarking sets for molecular docking. *J. Med. Chem.* 49, 6789-6801.

(61) Fukuzawa, K., Mochizuki, Y., Tanaka, S., Kitaura, K. Nakano, T. 2006. Molecular Interactions between Estrogen Receptor and Its Ligand Studied by the ab Initio Fragment Molecule Orbital Method. *J. Phys. ChemB.* 110, 16102-16110.

(62) Tanenbaum, D.M., Wang, Y., Shawn, W.P., Sigler, P.B. 1998. Crystallographic comparison of the estrogen and progesterone receptor's ligand binding domains. *Proc. Natl. Acad. Sci.* 95, 5998-600.

(63) Wang, X., Wang, Y., Chen, J., Ma, Y., Zhou, J., Fu, Z. 2012. Computational toxicological investigation on the mechanism and pathways of xenobiotics metabolized by cytochrome P450: a case of BDE-47. *Environmental Science & Technology 46*, 5126–33.

(64) Rydberg, P., Olsen, L. 2012. Ligand-based site of metabolism prediction for Cytochrome P450 2D6. J. *Med. Chem.* 3, 69-73.

(65) Sabo-Attwood, T., Kroll, K.J., Denslow, N.D. 2004. Differential expression of largmouth bass (*Micropterus salmoides*) estrogen receptor isotypes alpha, beta, and gamma estradiol. *Mol. Cell Endocrinol*. 218, 107-118.

CHAPTER 3.1. A SURVEY OF ASPARTATE-PHENYLALANINE AND GLUTAMATE-PHENYLALANINE INTERACTIONS IN THE PROTEIN DATA BANK: SEARCHING FOR ANION-Π PAIRS

A version of this chapter was originally published by Vivek Philip, Jason B. Harris, Rachel Adams, Don Nguyen, Jeremy Spiers, Elizabeth E. Howell, Robert J. Hinde, and Jerome Baudry:

Vivek Philip, Jason Harris, Rachel Adams, Don Nguyen, Jeremy Spiers, Jerome Baudry, Elizabeth E. Howell, Robert J. Hinde. "A Survey of Aspartate- Phenylalanine and Glutamate-Phenylalanine Interactions in the Protein Data Bank: Searching for Anion-π Pairs". *Biochemistry* (2011). 50(14):2939-50.

The work and writing in this chapter was contributed to by all authors. J. Baudry, E. Howell, and R.J. Hinde served as faculty advisors for all work presented in this chapter. J.B. Harris wrote code, generated all data, and constructed manuscript figures. V. Philip served as a senior student mentor, code writer, and code designer. R. Adams and D. Nguyen contributed legacy pieces of code. J. Spiers conducted the Figure 3.1.9 calculations.

Abstract

Protein structures are stabilized using non-covalent interactions. In addition to the traditional non-covalent interactions, newer types of interactions are suggested to be present in proteins. One such interaction, an anion-π pair, has been previously proposed where the positively charged edge of an aromatic ring interacts with an anion, forming a favorable anion-quadrupole interaction (Jackson et al., *J Phys Chem B* **2007**, *111*, 8242-8249). To study the role of anion-π interactions in stabilizing protein structure, pairwise interactions between phenylalanine (Phe) with the anionic amino acids, aspartate (Asp) or glutamate (Glu), were analyzed. Particular emphasis was focused on identification of Phe and Asp or Glu pairs separated by less than 7Å in the high resolution, non-redundant Protein Data Bank. Simplifying Phe to benzene and Asp or Glu to formate molecules facilitated *in silico* analysis of the pairs. Kitaura-Morokuma energy calculations were performed on roughly 19,000 benzene-formate pairs and the resulting energies analyzed as a function of distance and angle. Edgewise interactions typically produced strongly stabilizing interaction energies (-2 to -7.3 kcal/mol), while interactions involving the ring face resulted in weakly stabilizing to repulsive interaction energies. The strongest, most stabilizing interactions were identified as preferentially occurring in buried residues. Anion-π pairs are found throughout protein structures, in helices as well as β-strands. Numerous pairs also had nearby cation-π interactions as well as potential π-π stacking. While over a thousand structures did not contain an anion-π pair, the remaining 3134 structures contained approximately 2.6 anion-π pairs per protein, suggesting it is a reasonably common motif that could contribute to overall structural stability of a protein.

Introduction

Analysis of protein structure and ligand binding has been traditionally understood based on hydrogen bonds, van der Waals interactions, hydrophobic interactions, and ion pairs. However, other types of non-bonded interactions have been suggested to play a role in the stabilization of protein structures and of protein-ligand interactions. For example, hydrogen bonds involving CH as a donor group and π systems as acceptor

groups have been described, including formation of CH-π pairs[1,2] as well as CH to O pairs[3,4], where the aliphatic hydrogen forms weak bonds with nearby aromatic and carbonyl groups respectively. Other groups have proposed S-π bonds between cysteine and aromatic amino acids[5].

An additional type of non-bonded interaction includes the cation-π pair between aromatic sidechains (e.g. phenylalanine, tyrosine and tryptophan) and positively charged sidechains (lysine and arginine), where the cation interacts with the π electron cloud on the face of the aromatic ring. These cation-π pairs have been proposed to stabilize protein structures by ~3 ± 1.5 kcal/mol[6,7] Recently, an n to π* interaction[8] has been proposed to exist whereby a delocalized, lone pair of electrons from a backbone carbonyl atom interacts with the antibonding orbital of the next carbonyl oxygen. This interaction is predicted to provide up to 0.5-1.3 kcal/mol of stability to helices when it occurs in protein structures. Additional calculations suggest that π-π stacking is likely important[9-11]. While all these interactions are weak, numerous such occurrences can produce a substantial stabilizing effect on protein structure.

Lastly, we have proposed that an anion-π interaction exists which can contribute energetically to protein stabilization, ligand binding or protein-protein association[12]. In this present work, we focus on the prevalence of anion-π interactions in protein structures where negatively charged amino acids (aspartate or glutamate) can form energetically favorable pairs utilizing the positively charged ring edge of aromatic groups. The positive charge on the aromatic ring edge arises from the quadrupole moment of the sidechain, leading to the anion-quadrupole or anion-π name. As the description suggests, this interaction is related to the cation-π pair; however anion-π pairs facilitate an interaction between an anion and the aromatic ring edge rather than a cation with the ring face. Other groups have studied anion-π interactions in small molecules and have focused on electron-deficient π rings by incorporating strong electron-withdrawing substituents such as fluorobenzene derivatives, fluoro-*s*-triazine, and tetrafluoroethene as well as other aromatic molecules[13-25]. Experimental evidence for the existence of anion-π interactions in small molecules includes spectroscopic, NMR and crystallographic data of anion binding sites in electron deficient aromatics and host-guest molecular complexes, as well as other compounds found by screening the Cambridge Structural Database[26-30].

Early examples of anion-π interactions in biology appear to have been noticed, but not identified as such. For example, oxygen atoms and cysteines display a high propensity to occur at the ring edge of aromatic amino acids[31-34]. Additionally, the Atlas of Protein Sidechain Interactions indicates a statistical preference for an edgewise interaction between aromatic amino acids and Asp or Glu residues[35.] Studies by Kallenbach et al.[36-38] using short α-helical peptides with glutamate-phenylalanine pairs positioned at *i* and *i*+4 spacing indicate this pairwise interaction provides ~ 0.5 kcal/mol additional stability to the helix. For intermolecular binding interactions, Jouglin et al. found that tryptophan and tyrosine (as well as histidine and arginine) residues show the most enrichment at phosphoresidue binding sites [39]. Soga et al. propose aromatic

residues are “binding site-philic” as phenylalanine, histidine and tryptophan are most commonly found in binding of druglike molecules [40]. Most recently, studies of a limited four amino acid code employed at antibody binding sites find tyrosine as the dominant residue which provides tight binding to a host of antigens [41-43]. For the latter two examples, aromatic groups can play many roles during binding as they are amphipathic, can provide cation-π and anion-π interactions and the tyrosine and tryptophan sidechains can provide H-bonds. The relevance of the anion-π interaction is further recognized as important in more recent experimental as well as theoretical studies[11,44-49] and references therein.

Our present statistical analysis expands our previous theoretical study by searching for anion-π pairs in a nonredundant, high resolution subset of the Protein Data Bank (PDB). We focus on contributions involving phenylalanine as the aromatic partner and either aspartate or glutamate as the anion. First, angles and distances are calculated between these partners identified from the PDB screening. Then, using simplified models for each chemical group, interaction energies are obtained. We find a substantial number of such anion-π pairs to be present in the PDB with stabilizing energetics (more negative than -2 kcal/mol).

Materials and Methods

This section describes the steps taken to calculate energies associated with potential anion-π interactions in proteins. Only interactions between phenylalanine, simplified to a benzene ring, and aspartic acid or glutamic acid, simplified to formate, were studied. While tyrosine and tryptophan are aromatic amino acids which can be simplified to phenol and indole, both molecules contain dipole moments and can participate in hydrogen bond formation. Therefore, it is difficult to parse out the role of the anion-π interaction in their energetics, and they were not analyzed. Figure 3.1.1 shows a general flowchart of the steps in our analysis of biological anion-π interactions. The first step used a C++ program named STAAR (STatistical Analysis of Aromatic Rings)[12] to analyze the Hobohm and Sander subset[50,51] of the PDB, which includes non-redundant, high resolution structures. In our approach, only crystal structures with a resolution of $\leq$ 2 Å were analyzed; this corresponded to 4491 entries (March 2006 release). STAAR locates phenylalanine rings and determines their centers of mass (CM). For each aromatic ring, STAAR then calculates the distance *r* between the ring’s center of mass and the nearest oxygen atom in a Glu or Asp carboxylate group, as well as the angle θ between the plane of the ring and the vector connecting the ring center of mass with this oxygen atom. We adopted the cutoff criteria of Gallivan and Dougherty[7], i.e. those pairs possessing a distance of $r \leq 7$ Å were chosen for analysis to eliminate cases in which a water molecule could fit between the two residues and diminish the interaction energy. While STAAR can identify intermolecular pairs, in this study, we focus only on intramolecular pairs.

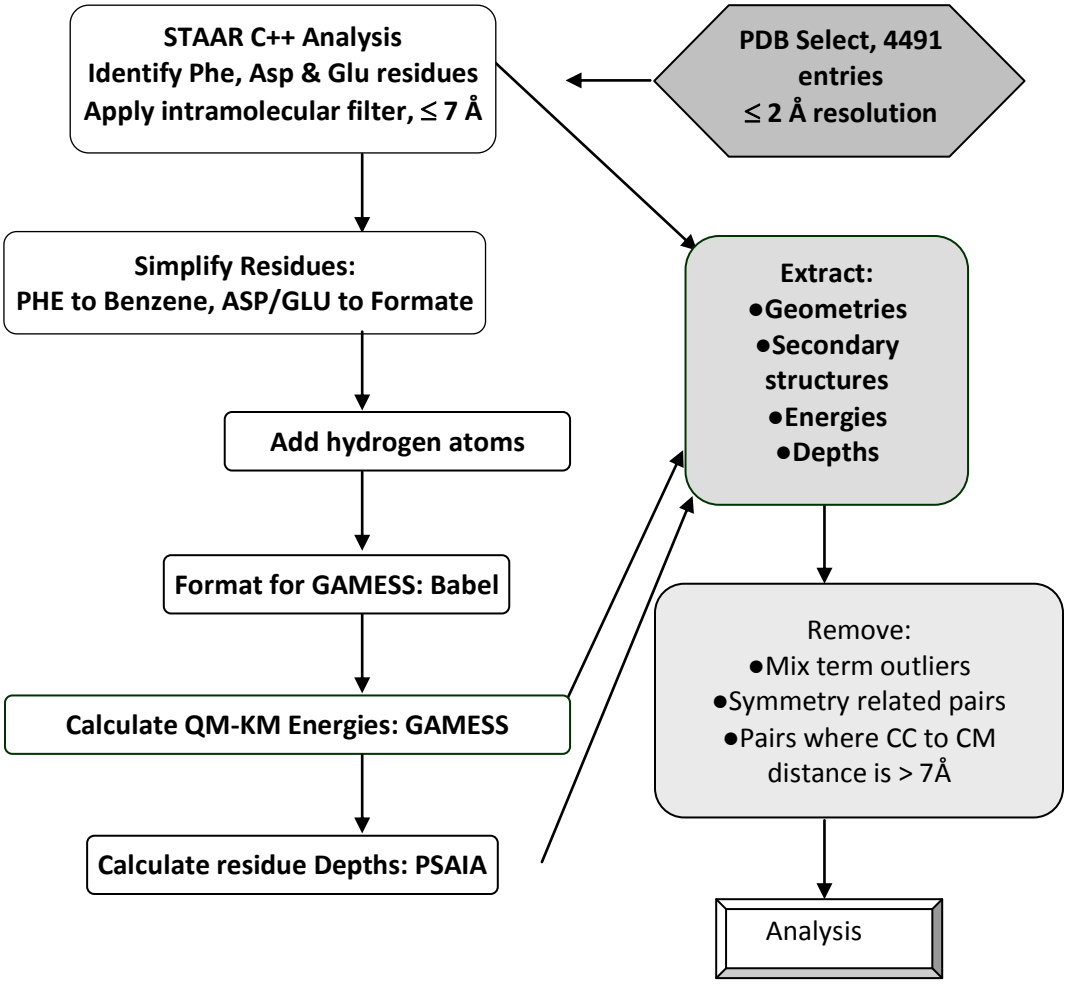


Figure 3.1.1 A flowchart of the calculations. CC describes the center of charge of the formate and CM describes the center of mass of the benzene molecule.

The next steps in our analysis simplified the Phe and Asp or Glu pairs to benzene -formate (BF) pairs. This was followed by addition of hydrogens to the BF pairs using ProDrg2 at http://davapc1.bioch.dundee.ac.uk/prodrg/[52]. Since the pK_a values for Asp and Glu are low (i.e., 3.5-4.5,)[53], we assume Asp and Glu are always ionized. The resulting file was converted from PDB coordinates to an xyz format using BABELwin[54]. PC GAMESS (June 1999 version) [55] was used to calculate the pairwise energies. Perl scripts and Excel spreadsheets were used to sort through the data. Later calculations used a Perl script to calculate the center of charge (CC) for formate. This allowed us to redefine *r* and θ as the distance and angle between the formate center of charge and the benzene center of mass.

An important step in our study of anion-π interactions was to identify a tractable, but energetically accurate calculation. Using Gaussian 03[56], we previously performed

quantum chemical calculations, corrected for basis set superposition error (BSSE), for optimized benzene-formate pairs BF1-BF4 (see Figure 3.1.2) at both the Hartree-Fock (HF) and second order Møller-Plesset (MP2) levels of theory[12]. To screen a large number of pairs, we hoped to identify an approach that would be more rapid than these first-principles quantum chemical calculations. We therefore looked for a linear correlation between the HF energies of the BF1-4 pairs with energies calculated using several semi-empirical approaches. Although Gallivan and Dougherty found the OPLS (optimized potentials for liquid simulations) forcefield worked well for identifying cation-π interactions[7], we did not observe a linear correlation between OPLS and HF energies. While reasonable correlations were observed using AM1 and PM3 forcefields, we ultimately used PC GAMESS running a Kitaura-Morokuma (KM) energy decomposition analysis[57-60]. Using the optimized BF1-BF4 pairs, a reasonably linear correlation (slope = 0.92, χ^2 = 0.988) was observed when the HF and KM energies were compared. To compare these two treatments for benzene-formate pairs derived from actual Phe with Asp or Glu residues that occur in the PDB, we calculated both KM and HF energies for 322 different pairs, as shown in Figure 3.1.3. The aug-cc-pVTZ atom-centered basis set was employed in all calculations[61,62]; spherical Gaussians were used except where otherwise indicated. Interaction energy calculations were performed with the benzene center of mass held at the origin of the spherical coordinate system and the benzene molecule stationed in the (*x*, *y*) plane, with CH bonds oriented along the positive and negative *y* axes. The counterpoise method[63] was used to correct for basis set superposition error in all HF benzene-formate calculations.

CHARMM22 Calculations 140 pairs from the PDB that were found to exhibit negative (i.e. attractive) anion-π *ab initio* interaction energies, were chosen to be analyzed by the CHARMM22 force field[64] implemented in the program MOE version 2009.10 (Chemical Computing Group, Ltd. Montréal, Canada). Calculations were performed in the gas phase with no cutoff value for non-bonded interactions. For each of the 140 pairs, the residues were isolated from their protein context and backbone atoms were deleted with the exception of the α-carbon, i.e. only side-chain and α-carbon atoms were considered. Hydrogen atoms were added and an energy minimization was performed on the hydrogen atoms only (heavy atoms were held fixed). Interaction energies were calculated by subtracting the sum of the CHARMM22 potential energies of the two individual amino acids from the CHARMM22 potential energy of the amino acid pair. The van der Waals and electrostatic contributions to the non-bonded interaction energies were also recorded. In another set of calculations, sidechain atoms were deleted from the models until only the functional groups remained. The potential interaction energy between these functional groups was calculated as described for the side chains for all the 140 anion-π pairs. Correlation coefficients were calculated between the energies derived from CHARMM22 and those interaction energies obtained from HF, MP2 and KM calculations for both side chain and functional groups.

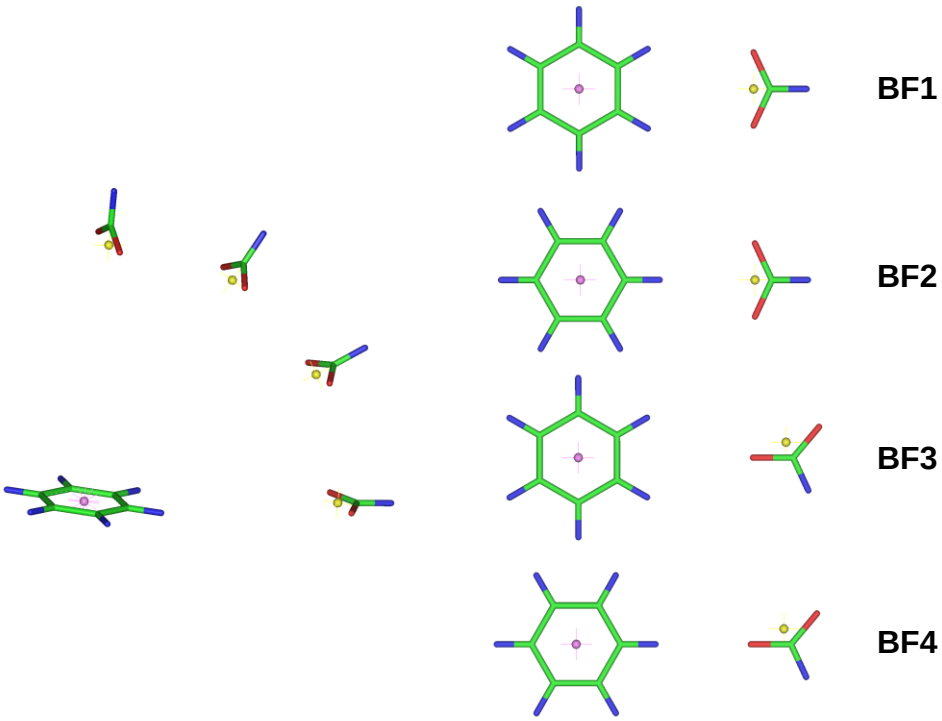


Figure 3.1.2. Benzene-formate pairs. Right: Four different coplanar dimers of benzene and formate were constructed and labeled BF1-4[12]. Left: The dihedral angle θ (between the planes containing the benzene and formate monomers) of the BF1 pair was varied between 0° to 90° by 30° increments. The center-of-mass position for benzene is shown in purple while the center-of-charge for formate is yellow. The atoms are colored as follows: carbon (green), oxygen (red), and hydrogen (blue).

Results

Screening the PDB A C++ program, STAAR[12], was used to identify Phe, Asp and Glu residues and to calculate the distance and angle between the aromatic ring and the carboxylate moiety. This was the first step in investigating the energetics associated with aromatic-anionic pair formation. Our approach follows the general procedure of Gallivan and Dougherty, who examined cation-π interactions in the PDB[7]. Their rationale for including an energy calculation step was based on both the complex electrostatic potential surfaces of aromatic residues as well as the inability of geometric criteria to differentiate between attractive and repulsive interactions.

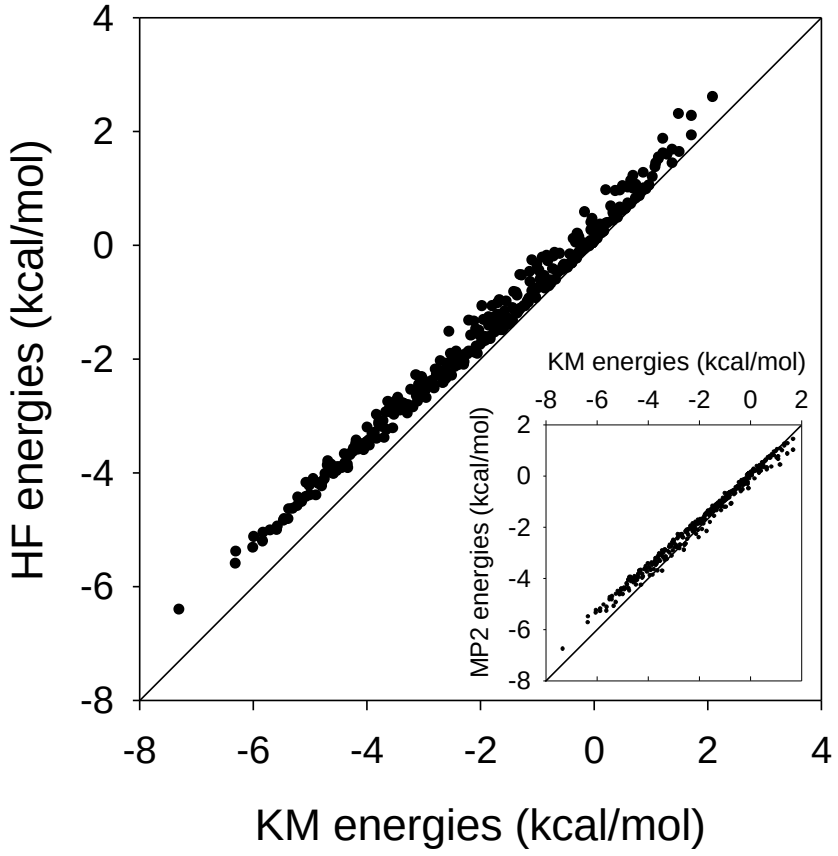


Figure 3.1.3. Correlation between Hartree-Fock (HF) and Kitaura-Morokuma energies. From over 4000 PDB files, greater than 19,000 phenylalanine with Asp or Glu pairs were identified and simplified to benzene-formate pairs. PC GAMESS was used to calculate interaction energies using a Kitaura-Morokuma (KM) energy decomposition approach. The pairs were sorted according to their interaction energies and approximately every fiftieth pair was selected for HF and MP2 energy calculations using a BSSE correction error[63]. The HF and KM energies from these 322 pairs are shown.

The high resolution, non-redundant Hobohm and Sander subset of the PDB (March 2006 release) was screened using the program STAAR[50,51]. A distance filter was employed to identify those aromatic-carboxylate pairs which were separated by ≤ 7 Å. An initial search found approximately 40,000 pairs that satisfied this distance criterion; however, because STAAR recognizes identical pairs from different, but symmetrical subunits, subsets of the identified pairs were redundant. For those PDB files that contained information regarding monomer identity (A,B,C, etc), a Perl script identified and removed the duplicates, retaining the pairs from only non-redundant monomers. After this step, approximately 22,000 pairs between Phe and either Asp or Glu in 3995 PDB files remained.

Next, the amino acid sidechains were simplified to benzene and formate; hydrogens were added using ProDrg2[52]. As PC GAMESS requires an xyz file format, the PDB coordinates were reformatted using BABELwin[65]. PC GAMESS[55] was then used to calculate the energies associated with pair formation using the Kitaura-Morokuma energy decomposition analysis[57-60]. We previously found this treatment provided good estimates of the interaction energy at low angles between the plane of the ring and the carboxylate, but did not accurately reflect the energetic contributions at higher angles, particularly at $\theta = 90^{o}$.

The resulting pairs were analyzed by various criteria. The KM calculation deconvolutes the total energy into electrostatic (ES), polarization (PL), charge transfer (CT), exchange repulsion (EX) and mix terms. The "mix" term balances any differences that arise between the sum of the ES, PL, EX and CT terms with the total interaction energy. Supplemental Figure 3.1.S1 plots the total energy versus the "mix" value. As the bulk of "mix" values occur between -0.25 and 0.25 kcal/mol, approximately 500 outliers with mix values > |0.25| kcal/mol were removed from further analysis. This step was performed because when the mix term is large in magnitude, it is difficult to interpret the physical significance of the other energy terms.

The distance *r* calculated using STAAR is the distance from the center of mass of benzene to the closest oxygen in the carboxylate group. Using a Perl script, the position for the formate center of charge (CC) was calculated, allowing recalculation of the *r* value as the distance between the formate CC and the benzene center of mass. After this step, another distance filter step allowed removal of approximately 5000 pairs with r > 7Å. A final Perl script counted the number of BF pairs per PDB file. Three files (1A9X, 1E6Y and 1K8K) exhibited >50 occurrences of anion-π pairs. Upon inspection of the PDB file, symmetry- related monomers were found to be present in 1A9X and 1E6Y, although not specified by a chain ID. These were manually removed, leaving 17,042 pairs for analysis.

Angle and Distance Analysis From our previous QM calculations on optimized BF pairs[12], we found that the strongest interaction energies were associated with edgewise interactions. This pattern arises from the positive electrostatic potential at the ring edge compared to a negative electrostatic potential at the ring face associated with the π electron clouds. The coplanar or edgewise interaction trend continues to be observed in crystal structures from the PDB, as shown in Figure 3.1.4 where a stacked bar plot displays the number of intramolecular BF pairs compared to the total interaction energy for increasing values of θ in 10^{o} increments. The total number of pairs is noted for each bin of θ values. Supplemental Figure 3.1.S2 compares the fraction of pairs residing in a particular 10-degree bin with the fraction that would be observed for a random, uniform distribution of BF separation vectors. We see that the number of pairs increases as θ decreases, and that more pairs are observed at small θ values (and fewer pairs at large θ values) than would be expected from a uniform distribution. Similar trends were observed in our earlier study[12] of a smaller ensemble of proteins from the PDB.

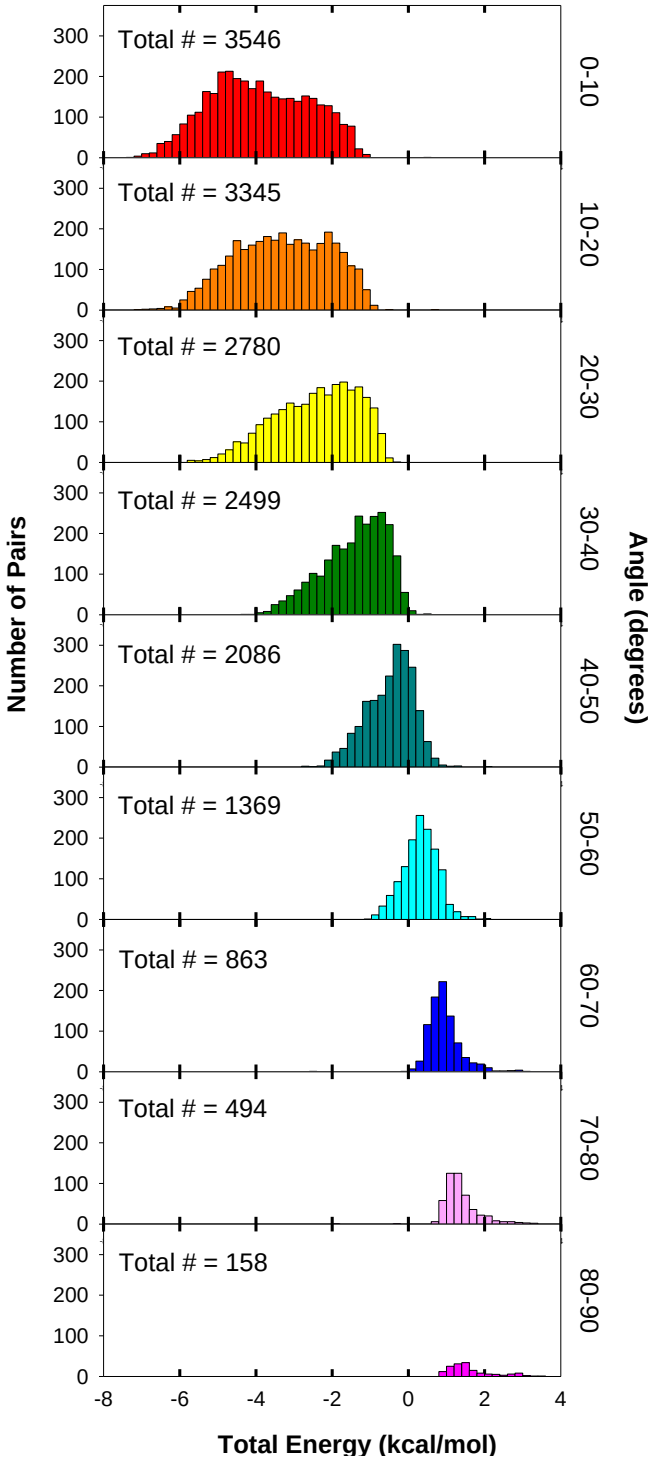


Figure 4. A histogram analysis of the interaction energies as a function of angle. Interaction energies for the BF pairs derived from the PDB were binned according to their dihedral angle θ and plotted. As the angle increases, the number of pairs decreases as does the energy.

To focus on those pairs with stabilizing interaction energies from -8 to -2 kcal/mol, the data were yet again filtered, leaving 8260 pairs in 3134 PDB files. Only 6 pairs were identified with energies more negative than -7 kcal/mol. The strongest attractive interaction (-7.27 kcal/mol) arises for the Phe97-Asp23 pair in the lignin peroxidase structure (1LLP). Most pairs with energies more negative than -7 kcal/mol resemble the BF1 configuration, with slight variations in the distance and angle, θ. The energies calculated for many of the anion-π interactions are substantially stabilizing, with roughly 39% of this 8260 pair subset having predicted interaction energies of $\leq$ -4.0 kcal/mol.

As the radius of a sphere increases, so does the volume. Assuming a random distribution, one would expect increasing numbers of pairs to be found as the radius between the benzene center of mass and the carboxylate center of charge increases[7,35]. This is seen in Figure 3.1.5 (black bars) for the 17,042 BF pairs encompassing the -8 to +5.6 kcal/mol energy range. However, when the 8260 pairs with energies between -8 and -2 kcal/mol are analyzed, there are fewer pairs at longer distances (gray bars in Figure 3.1.5), consistent with closer contact correlating with more negative interaction energies.

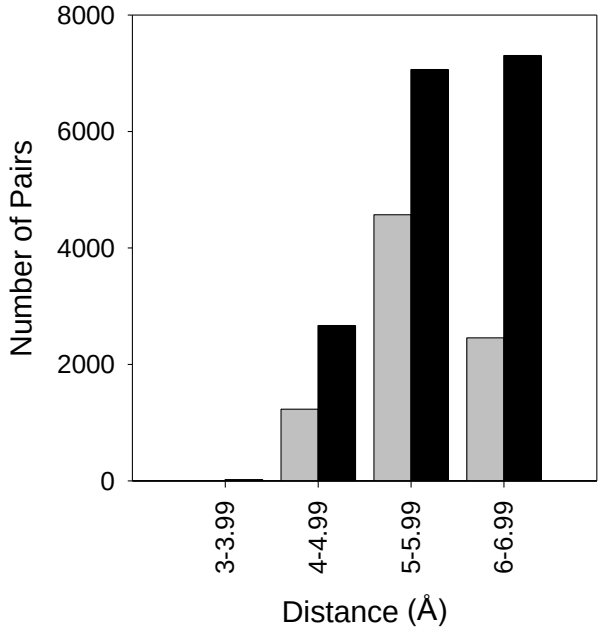


Figure 3.1.5. Distance relationships of the BF pairs. The gray bars describe the number of pairs possessing interaction energies from -8 to -2 kcal/mol as a function of distance. The black bars provide a count of the number of total pairs for these distances for all the BF pairs in the -8 to +2 kcal/mol energy range.

The strongest attractive interactions for the BF1 and BF2 pairs occurred with an approximately 4Å separation distance (see Figure 2 of reference *(12)*)[12] while the optimal energies for the BF3 and BF4 pairs occurred at roughly 5.5 Å. When the separation distances between the Phe and Asp or Glu pairs identified in this study are considered (Figure 3.1.5), the largest number of pairs occurs between 5-6 Å. This suggests that either the pairs found in nature mimic the BF3 and BF4 pairs or that the protein context plays a role in residue positioning, allowing other configurations of benzene and formate. To determine if a large population of BF3-BF4 pairs are present in the dataset, we noted that the distance between oxygen atom 1 (oxy1) of formate and the benzene CM should be similar to the distance between the oxygen atom 2 (oxy2) of formate and the benzene CM for the BF1 and BF2 pairs. In contrast, the difference in these distances for the BF3 and 4 pairs should be larger. Therefore we calculated the absolute value of the difference of these distances and plotted these values as a function of energy. We did not observe a bimodal distribution as might be expected if the pairs segregated into BF1-BF2 like and BF3-BF4-like groups; rather the pairs sample a large range of sequence space (see supplemental Figure 3.1.S3). This analysis indicates proteins show more structural variation than the optimized BF pairs initially considered.

Structural Analysis Figure 3.1.6A indicates most anion-π pairs are widely separated in the primary sequence of their respective proteins. However, 8% of the total pairs occur next to each other with an i and i+1 spacing. Additionally, 11% show an i and i+2 spacing, consistent with β-sheet interactions. Finally, 14% show an i and i+3 or an i and i+4 pairing, consistent with residues interacting in an α-helix.

The secondary structure information of the proteins exhibiting anion-π pairs was extracted for a subset of 6934 pairs using a Perl script. Figure 3.1.6B shows interactions involving an undesignated structure are most common, followed by helix-helix interactions. Strand-strand and helix-strand interactions also occur. As Phe, Glu and Asp have strong propensities to be in helices[66,67], these results are not surprising. Additionally, we note Phe has a strong propensity to be found in a β-sheet, however Asp and Glu do not[68-70].

Since the periodicity of a β-strand is 2, it was surprising to observe β-strands in the i and i+1 as well as i and i+3 categories. For the i and i+1 pairs, the residues often occur on the surface of the protein and near the end of a strand. The sidechain of the anionic residue folds back on top of the Phe residue. Figure 3.1.6C shows an example of one such pair in 1VRM. This could potentially be of importance in "edge protection strategies" preventing amyloid formation[71].

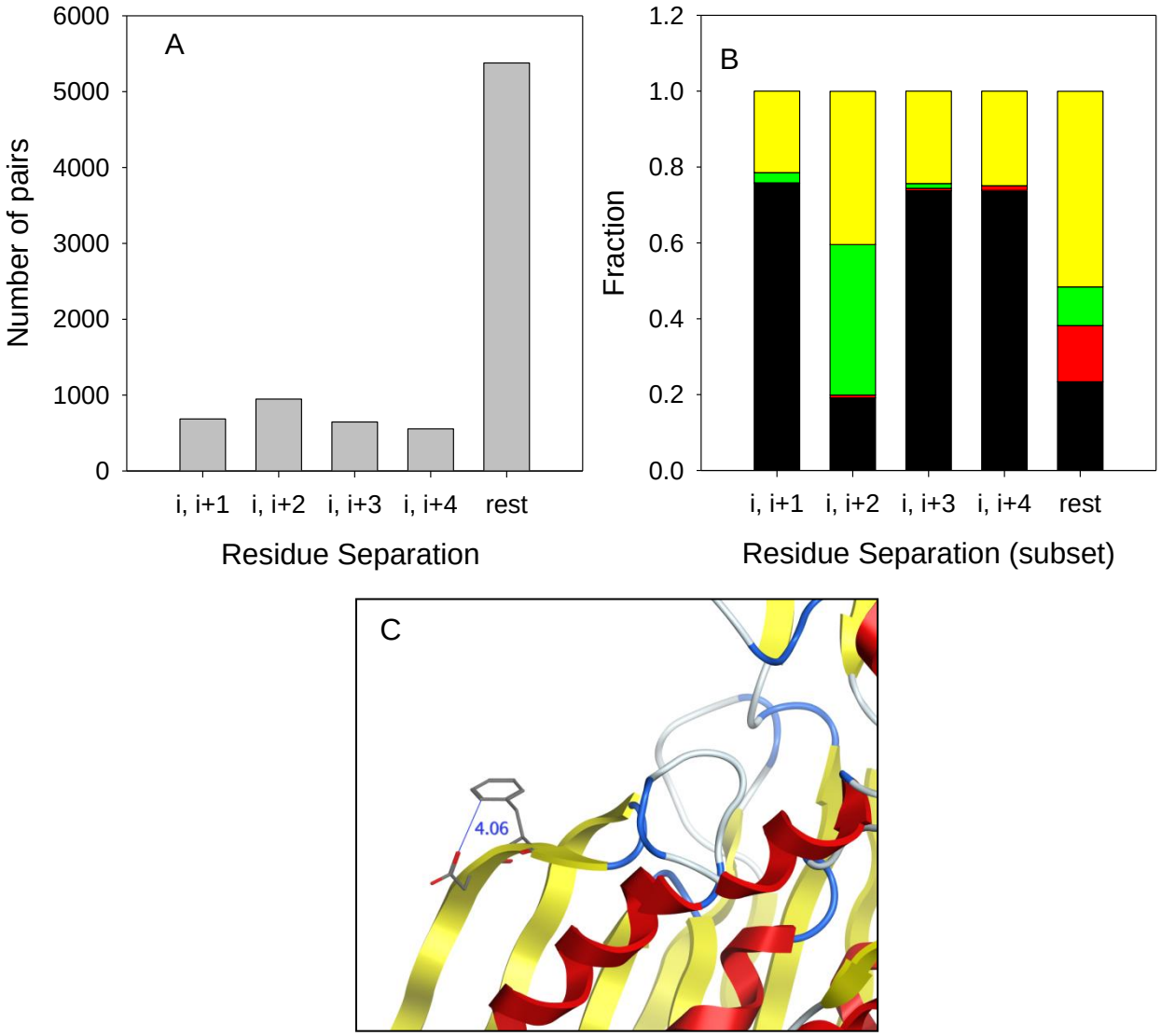


Figure 3.1.6. Structural analysis of 8260 BF pairs possessing energies from -8 to -2 kcal/mol. In panel A, the primary sequence separation of Phe and Asp or Glu residues in the BF pair is shown. A separation of greater than 4 amino acids is preferred. Separations of i and i+2 may indicate placement of the BF pair in a β-strand, while separations of i and i+3 or i+4 may indicate the BF pair occurs in an α-helix. We were able to extract the secondary structure designation from a subset of 6934 pairs. Fractional preferences for these BF pairs are presented in panel B. Black denotes α-helix-α-helix interactions, green describes β-strand-β-strand interactions, red depicts α-helix-β-strand interactions while yellow indicates one or both of the structural elements is not designated. Panel C shows the structure of one β-strand-β-strand i and i+1 interaction occurring in 1VRM between D50 and F51. The distance between the CD and OD1 atoms is 4.06 Å. Strands are shown in yellow, helices in red, turns in blue and undesignated structure in white.

The PSAIA program was used to address the question of whether the BF pairs are buried or whether they occur on the surface of the protein. This program calculates the average residue depth, defined as the distance in Å from the closest solvent accessible atom[72]. A value of 0 describes a fully accessible atom, while values greater than 0 describe buried atoms, with larger values describing more deeply buried atoms. Use of PSAIA allows automation of these calculations[73]. The average residue depths for each member of the pair were added and plotted against the total KM interaction energy, as shown in Supplemental Figure 3.1.S4. The points are randomly scattered and do not show a trend between depth of the roughly 17,000 BF pairs and their interaction energy. However as shown in Figure 3.1.7, a plot of the average residue depths for those 100 BF pairs possessing the most negative energies does show a preference for partial burial of the residues. As an electrostatic component of the anion-π interaction exists, burial of the pair would minimize potential disruption of the anion-π geometry by water[74]. Protection from solvent appears important as a recent computational analysis of cation-π interactions indicates weaker energies in the presence of water[75]. In addition, some site directed mutagenesis studies of surface cation-π interactions in four different proteins indicate that cation-π interactions are "at best weakly stabilizing and in some cases are clearly destabilizing" at room temperature[76].

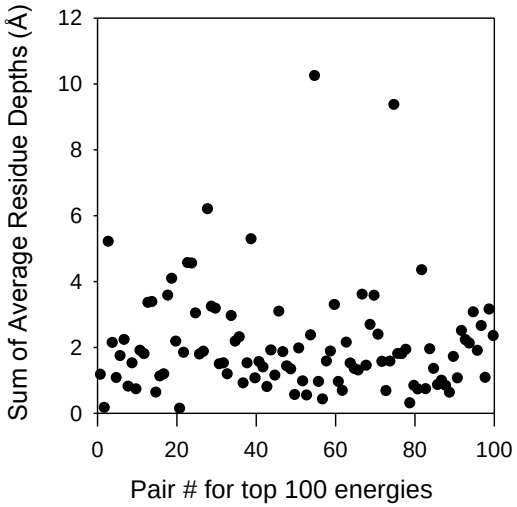


Figure 3.1.7. Depth of the top 100 BF protein pairs calculated by the PSAIAA program[72,73]. BF pairs were sorted according to their energies and the top 100 are presented, with pair number 1 having the most negative energy (-7.27 kcal/mol). The calculated energy for the 100th pair is -6.27 kcal/mol. The average residue depths for the Phe and Asp or Glu residues were summed and the sum plotted. An additional plot for the sum of the average residue depth for Phe and Asp or Glu for the entire data set of roughly 17,000 pairs was constructed and is shown in Supplemental Figure 3.1.S4. No pattern was observed in the larger dataset.

No apparent preference for Glu or Asp exists in their interaction with phenylalanine. The ratio of Asp to Glu involved in anion-π pairs is very close to the ratio of Asp to Glu residues observed in our entire protein structure sample.

How often do BF pairs occur? We did not identify energetically significant pairs (-8 to -2 kcal/mol range) in 1357 PDB files. For the remaining 3134 PDB files, 8260 BF pairs were found, corresponding to approximately 2.6 pairs per structure. For comparison, Gallivan and Dougherty found an average of 1 energetically significant cation-π interaction per 77 residues in a protein[7].

Interesting Clusters Clusters of non-bonded interactions (involving greater than one anion-π pair) were found in the protein structures. Figure 3.1.8 shows several potentially interesting arrangements. For example, panels A and B show the presence of several aromatic groups surrounding one anion. This configuration was reasonably common as 365 clusters were identified with two phenylalanines interacting with one Glu or Asp. Panel B gives an example where π - π stacking may occur between 2 phenylalanines interacting with 1 Asp. Panel C shows an example of a cation-π interaction occurring concurrently with an anion-π interaction and panel D shows a potential network of 4 BF pairs in 1EP0. Panel F shows that often several anions may cluster around an aromatic group. Since anion-π pairs utilize polarization as an important component of their interaction energy, the stabilizing mechanism involved in this type of cluster is not clear, however it may be that nearby carboxylates have altered pK_a values such that some of these residues are protonated. We identified 620 BF pairs that contain one phenylalanine and two Glu or Asp residues. Panel G shows a histidine near a BF pair. While histidine is aromatic and could engage in stacking interactions, it also has the possibility of being protonated and participating in a cation-π interaction. The number of possible interactions expands if nearby Lys, Arg, His residues are considered. As support for the context of the anion-π interaction in the protein structure affecting the energetics of the system, we note recent quantum mechanical and crystallography studies of small molecules find synergistic effects between anion-π interactions and H-bonding networks, lone pair-π CH-π, cation-π and π-π stacking interactions[77-80]. The interplay between these various elements appears to alter the overall energetics, often in a synergistic manner.

Table 3.1.1. Correlation coefficients between calculated interaction energies with various quantum mechanical treatments and CHARMM22.

	HF	MP2	KM
Functional groups	0.73	0.82	0.75
Side chains [a]	0.61	0.68	0.65
HF		0.97	0.99
MP2			0.99

[a.] after deletion of two pairs of residues out of the initial 140 pairs, as described in results.

Figure 3.1.8. Interesting arrangements of anion-π pairs. Panel A shows a series of interactions between D145, F217, F218, and E274 residues in the 1AYX PDB file. In particular, E274 interacts with both phenylalanines. Panel B, interactions between D145, F254, and F283 in 1XSZ. Here, ring stacking between the phenylalanines appears to be occurring. Panel C shows the juxtapositioning of R42 near the F7 and E58 anion-π pair in 2FGQ. The R42 residue engages in a cation-π interaction with F7, as per the web-based Capture program (http://capture.caltech.edu/), which predicts potential cation-π interactions[7]. Polarization is predicted to contribute strongly to both anion-π and cation-π interactions, suggesting each type of interaction could enhance the other[7,12]. Panel D shows a long series of anion-π interactions in 1EP0, involving F4, F38, D84, E111, F112 and F122. Panel E depicts a cluster of three phenylalanines (F58, F62, and F297) interacting with D293 in 1VFL. Panel F illustrates the interactions between F173, D168, D113 and E161 in 1CPO. The rationale for the interaction of several carboxylates with one Phe is not as clear. It may be that a subset of the Asp or Glu residues are protonated. The atoms are colored as follows: carbon (gray), oxygen (red), and nitrogen (blue). Panel G shows the juxtapositioning of H399 near F708 and E749 in the 1T3T pdb file. Depending on its ionization state, H399 could potentially either engage in ring stacking with F708 or form a cation-π interaction with F708.

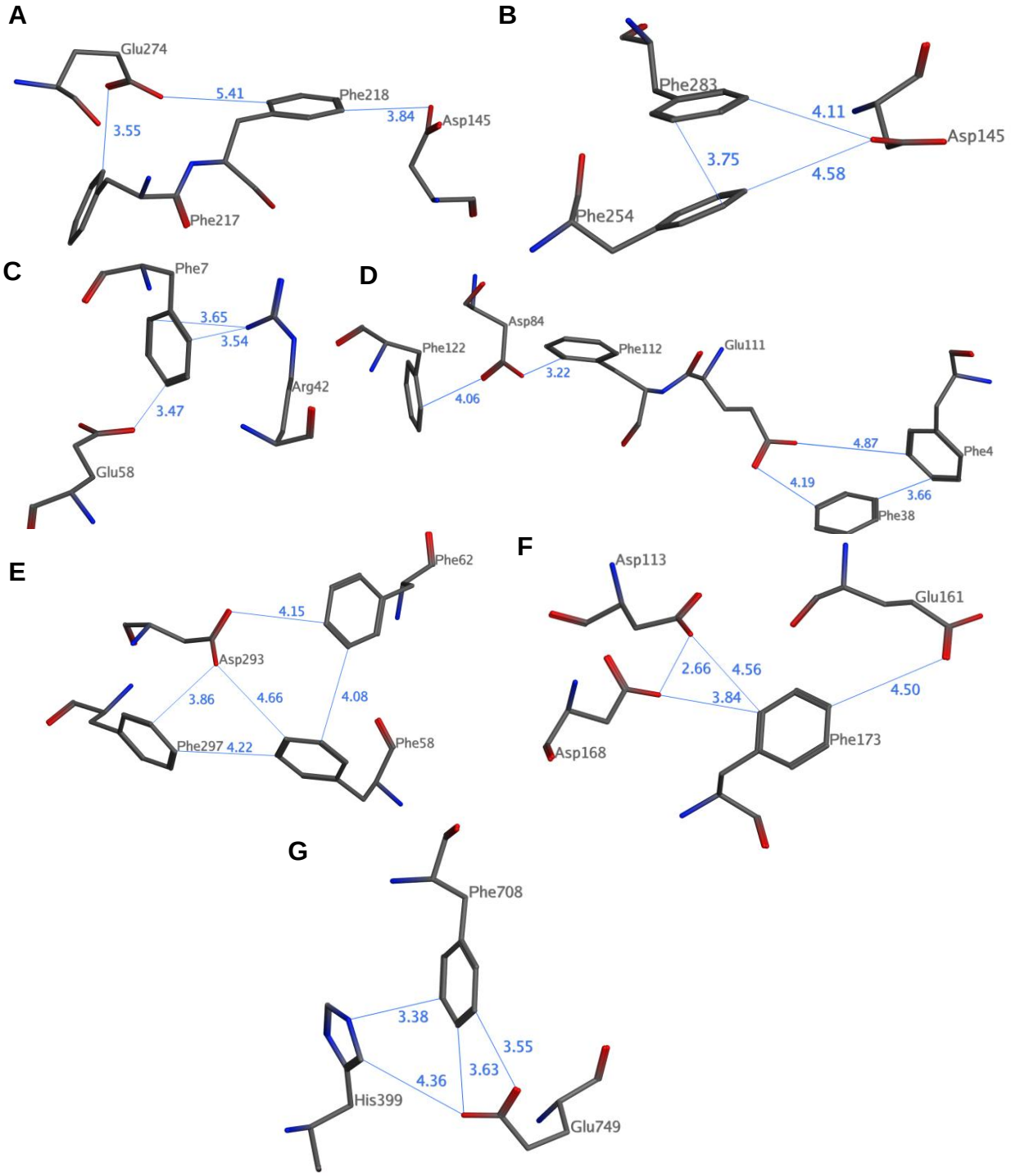


Figure 3.1.8 continued

CHARMM22 Calculations Figure 3.1.9 plots the CHARMM22 interaction energies for the BF functional groups vs. the energies for the MP2 calculations. The correlations between CHARMM22 and *ab initio* functional group interaction energies range from 0.73 to 0.82, depending on the level of theory considered. The correlation coefficients between empirical and *ab initio* interaction energies are given in Table 3.1.1. The highest correlation (0.82) is obtained when comparing force field calculations with the MP2 level of theory, suggesting that the attractive van der Waals interactions (implicitly included in force field parameterization, but not well represented in HF and MK calculations) significantly contribute to the correlation between classical and quantum-chemical models. As found previously[12], CHARMM22 interactions were under-evaluated when compared to MP2 interaction energies. No correlation was found between the angle and distance geometrical parameters between the interacting residues and the difference between MP2 and CHARMM22 interaction energies. This suggests that the difference between CHARMM22 and MP2 energies is systematic and independent of the geometry of the anion-π pairs.

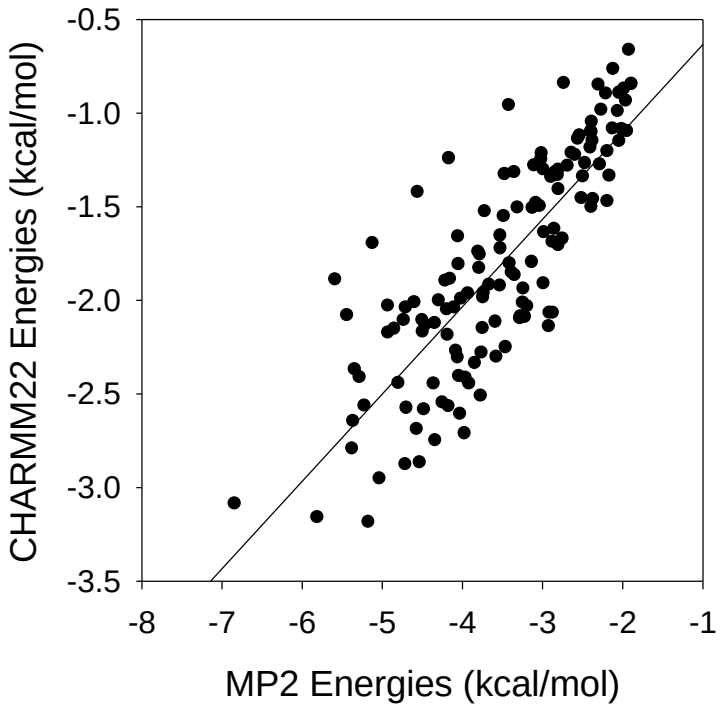


Figure 3.1.9. MP2 vs. CHARMM22 interaction energies, for functional groups. A subset of data from Figure 3.1.2 in the -8 to -2 kcal/mol energy range (138 BF pairs) were analyzed in CHARMM. A linear fit with a correlation coefficient of 0.82 was observed. The slope of the line is 0.47, indicating CHARMM can capture close to 50% of the MP2 interaction energy.

Correlation coefficients between *ab initio* and CHARMM interaction energies calculated for the entire side chains and alpha carbons are significantly lower than those for the functional groups only. The sidechain interaction energies range from 0.61 to 0.68, the highest correlation being between the empirical force field and MP2 level interaction energies. This was after the exclusion of two of the 140 pairs (Phe 337/Glu 295, from structure 1HLE; and Phe 388/Glu 285, from structure 1TQ4) that exhibited high repulsive interaction energies even after energy minimization of the hydrogen atoms. The highest correlation coefficient of 0.82 found here in the case of functional groups is close to the value of 0.89 obtained[81] when comparing CHARMM22 and MP2 interaction energies for 315 cation-π pairs identified from the Protein Data Bank[82]. It is not known whether the slightly lower correlation coefficient obtained in the present study originates from intrinsic parameterization differences that would surface when comparing cation-π or anion-π interactions by the CHARMM forcefield, or if it is only an effect of the lower number of pairs considered here. The correlation between CHARMM and *ab initio* interaction energies obtained from the entire side chain calculations (0.61 to 0.68), however, are significantly lower than that obtained in a cation-π study[81], which exhibited a correlation coefficient of 0.89 between CHARMM and MP2 interaction energies.

Discussion

The proposal that anion-π interactions exist may seem unexpected; because, at first consideration, this idea predicts an interaction between an electron-donating anion and an aromatic π cloud. However the quadrupole moment associated with aromatic groups results in points near the ring edge possessing a positive electrostatic potential, while points above and below the ring display a negative electrostatic potential. Anions can thus favorably interact with the ring edge[12].

Biological Relevance of the Anion-π Interaction Our present study identifies anion-π interactions between Phe and Asp or Glu as reasonably common in protein structures. Approximately two energetically favorable Phe-Asp or Phe-Glu pairs, with energies in the -8 to -2 kcal/mol range as calculated by the KM energy decomposition analysis, occur per PDB file. As shown in Figure 3.1.4, to achieve these interaction energies, an angle dependence is required, with a strong preference for edgewise interactions. Although favorable stacking interactions are possible when the Asp or Glu residue is positioned above the aromatic ring of the Phe residue[11], Supplemental Figure 3.1.S2 shows that in our database of structures, these geometries (with large θ values) are less common than would be expected from a uniform distribution.

The predicted energies describe gas-phase calculations. *In vivo*, water and other protein atoms will be present, which might be expected to screen the overall interaction energy. However a significant number of Phe-Asp or Phe-Glu pairs appear to be buried, as calculated by the PSAIA program (Figure 3.1.7 and Supplemental Figure 3.1.S4), which would minimize the screening of anion-π interactions for direct disruption of the anion-π pairs by water molecules.

The overall interaction energies may potentially be modulated by the environment of the Phe-Asp or Phe-Glu pair in the protein. For example, while we focus on Phe-Asp and Phe-Glu pairs, nearby tyrosines, tryptophans and neutral histidines could provide additional polarization, stacking effects, and/or H-bonds. Nearby arginines, lysines and protonated histidines could provide cation-π interactions that could enhance the anion-π interaction. Finally, other nearby anions and/or anion-π pairs could either perturb pK_a values or form a network of interactions (see "interesting clusters" in Figure 3.1.8). While the anion-π interaction may be weak, it is still significantly above k_BT (i.e., ~0.6 kcal/mol at 300K) and can occur frequently, rendering its overall effect on protein structure significant. Also, when a large network of interactions is considered, cooperativity could result, either enhancing or diminishing the overall effect.

How do the energetics of the anion-π and cation-π interactions compare? We might predict that the edgewise nature of the anion-π interaction, and the anisotropy of the polarizability tensor of aromatic rings, would favor the polarization term compared to that in the cation-π interaction. In addition, both the electrostatic and polarization effects in the anion-π interaction are most favorable for edgewise approach of the ion. In contrast, for cation-□ interactions, the electrostatic component of the interaction is most favorable for perpendicular geometries while the polarization component of the interaction is most favorable for edgewise approaches of the ion. These two elements suggest the energetics of anion-π interactions should be at least similar in scope to those of cation-π pairs. Using an OPLS forcefield, Gallivan and Dougherty found the average strength of a cation-π interaction involving Lys was -3.3 $\pm$ 1.5 kcal/mol and was -2.9 $\pm$ 1.4 kcal/mol when involving Arg[7]. Additionally, in studying the cation-π interaction with a simplified benzene-ammonium pair, Aschi et al. predicted an energy of -4.4 kcal/mol using a KM energy decomposition analysis. These values are clearly within the predicted range of the anion-π interaction shown in Figure 3.1.4.

We can expand our thinking about anion-π interactions to include protein-ligand binding. For example, numerous biologically relevant anions exist, such as DNA, RNA, phosphorylated proteins, ATP, NADPH, membrane bilayers, etc. If binding involves ion pair formation, binding specificity becomes an issue as other ions can compete. In contrast, neutral residues can provide both contacts, steric constraints and potentially greater ion selectivity by use of polarization during binding[83]. Both cation-π and anion-π interactions could participate in this fashion to facilitate both affinity and selectivity in binding of ions. Also of note, the desolvation penalty for anion-π and cation-π pair formation should not be as large as for ion pairs[84,85]. It will be interesting to extend the present study to protein-ligand complexes, and identify potential anion-π interactions between protein receptors and their ligands.

Computational Prediction Status The correlation between CHARMM22 and *ab initio* interaction energies (using the entire side chain) was 0.61-0.68, as shown in Figure 3.1.9. This result suggests that *ab initio* anion-π interaction energies are reproduced correctly by the CHARMM empirical force field for basic functional groups,

albeit with a magnitude that underestimates the *ab initio* results. Since *ab initio* or semi-empirical calculations have not been performed for the entire side chains for the anion-π pairs studied here, it is not possible to definitively quantify how well force field parameterization reproduces anion-quadrupole interactions for molecules larger than functional groups. However, these preliminary results suggest that the force field calculations are able to correctly assign a global ranking of the relative interaction energies between different pairs, but they underestimate the *ab initio* anion-π interaction energies.

While cation-π, anion-π, CH-π as well as other interactions appear important to protein structure formation, stability and dynamics, the impact of incorporating the present results on protein structure prediction remains unclear, in particular as to their potential influence on proteins' backbone structure and protein folding. Our previous theoretical studies of simplified phenylalanine and Glu or Asp pairs found the charge-quadrupole term contributes between 30 to 45% of the total MP2 energy, and the rest of the interaction energy arises mostly from polarization contributions[12]. However, most force-fields do not include explicit, "on-the-fly" polarization and multipole terms in the calculation of potential energies. While the global effect of polarization on the structure is included in the force field parameterization process, recent studies have either proposed or added polarization terms to forcefields[86,87]. These new force fields are expected to improve the agreement between empirical force field and *ab initio* results for anion-π interactions. It will be particularly interesting to see if an improved force-field treatment of the anion-π interactions will amplify the improving trend of protein structure predictions observed in the Critical Assessment of protein Structure Prediction (CASP) experiments[88] and references therein.

Acknowledgment

We thank Jordan Grubbs for construction of the images in Figure 3.1.8.

Abbreviations

HF, Hartree-Fock; MP2, second order Møller-Plesset energy calculations; BF, benzene-formate pairs; STAAR, a C++ computer program that provides STatistical Analysis of Aromatic Rings**;** PDB, protein data bank; OPLS, optimized potentials for liquid simulations; KM, a Kitaura-Morokuma energy decomposition analysis**;** ES, electrostatic term; PL, polarization term; CT, charge transfer term; EX, exchange repulsion term; QM, quantum mechanical; MM, molecular mechanics treatment; CC, center of charge; CM, center of mass; STAAR, a C++ program named STatistical Analysis of Aromatic Rings; BSSE, basis set superposition error.

Supplemental Figures

Four supplemental figures are included. The first plots the KM mix term versus the total energy. The second figure plots the observed fractional occurrence of benzene-formate pairs in the PDB as a function of the angle θ compared to the expected occurrence calculated from volume. The third figure plots distance differences vs. total interaction energy where the distance describes the difference between oxygen 1 (oxy1) and the benzene center of mass (CM) subtracted from the distance between oxygen 2 (oxy2) and the benzene CM and plotted as an absolute value. A final figure describes the average residue depth of the BF pairs in the protein structures calculated using the PSAIAA program versus the interaction energy. This material is available free of charge via the Internet at http://pubs.acs.org.

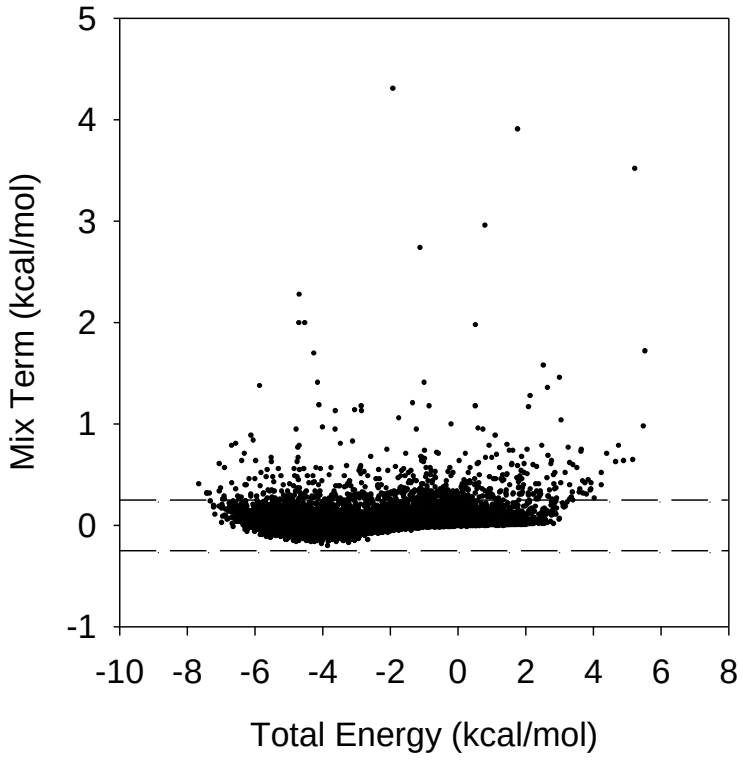


Figure 3.1.S1. Plot of total interaction energies vs. "mix" term values. The KM energy calculations include a "mix" term which describes the difference between the total interaction energy and the electrostatic, polarization, charge transfer and exchange repulsion terms. Dashed lines denote boundaries where mix terms outside the -0.25 to 0.25 kcal/mol range occur. 480 outliers were removed based on this analysis.

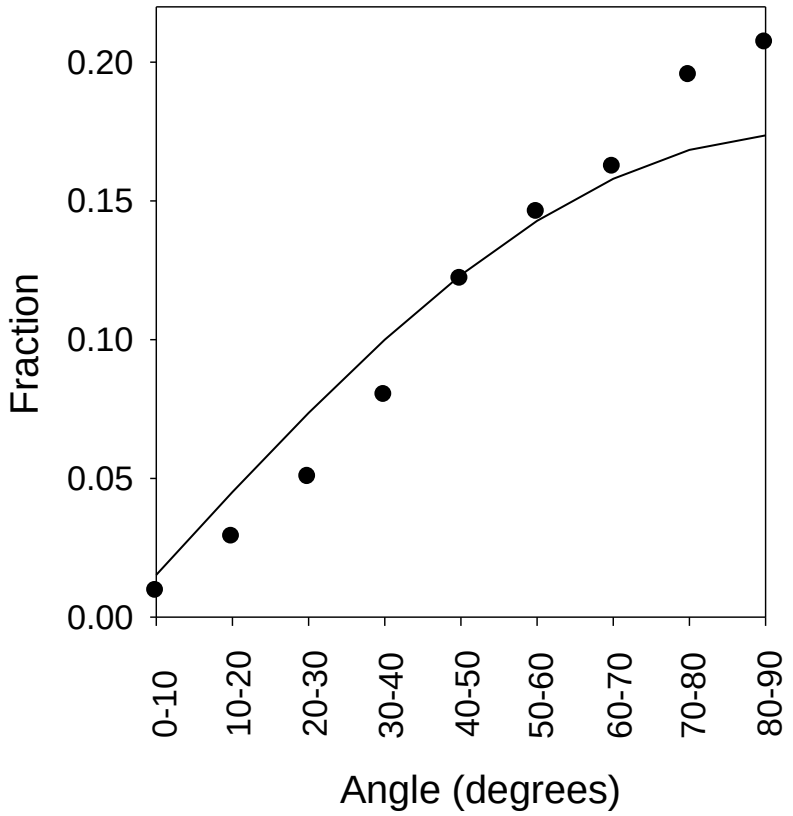


Figure 3.1.S2. The fractional occurrence of BF pairs as a function of θ. Theta was varied from 0-10°, 10.0001-20°, etc and the number of interacting pairs in the PDB Select determined using STAAR. The fraction of pairs that occurs in each 10° increment of θ was then calculated; these are circle points. The line describes the fractional number of pairs expected based on statistical considerations, specifically these values were calculated by computing the volume of the sector (above and below the spherical bisector) divided by the total spherical volume.

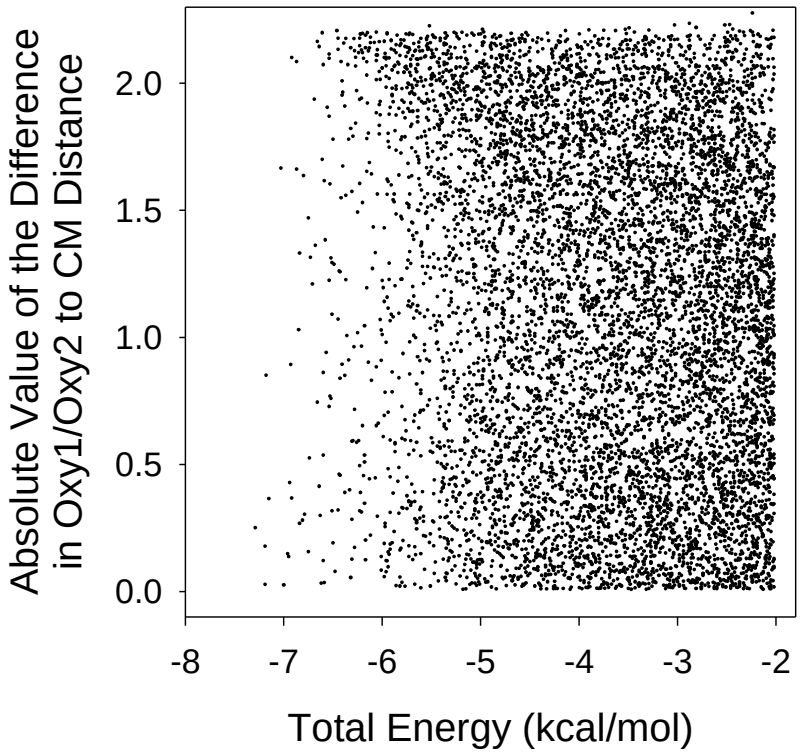


Figure 3.1.S3. Plot of distance differences vs. total interaction energy. The distance between oxygen 1 (oxy1) and the benzene center of mass (CM) was subtracted from the distance between oxygen 2 (oxy2) and the benzene CM and plotted as an absolute value. This distance difference should be small for BF1 and BF2 pairs (see Figure 3.1.2 in the main manuscript) and larger for the BF3-BF4 pairs. The data subset containing ~8200 pairs from -2 to -8 kcal/mol was analyzed.

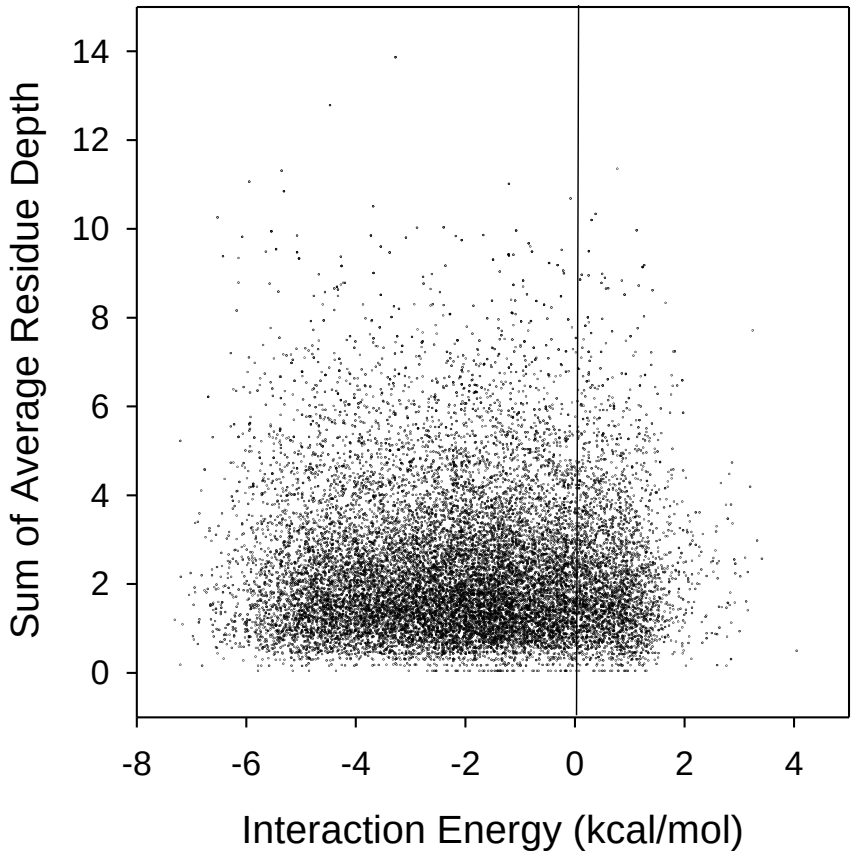


Figure 3.1.S4. Depth of BF pairs in the protein structures. The average residue depth was calculated using the PSAIAA program (18-19). The values for each member of the BF pair were summed and plotted vs. the Kitaura-Morokuma interaction energy. No overall pattern for the roughly 17,000 pairs was observed.

References

(1) Brandl, M., Weiss, M. S., Jabs, A., Suhnel, J., and Hilgenfeld, R. 2001. CH-pi interactions in proteins, *J. Mol. Biol. 307*, 357-377.

(2) Weiss, M. S., Brandl, M., Suhnel, J., Pal, D., and Hilgenfeld, R. 2001. More hydrogen bonds for the (structural) biologist, *Trends Biochem. Sci. 26*, 521-523.

(3) Jiang, L., and Lai, L. 2002. CH···O hydrogen bonds at protein-protein interfaces, *J. Biol. Chem. 277*, 37732-37740.

(4) Guo, H., Beahm, R. F., and Guo, H. 2004. Stabilization and destabilization of the Cd-H···O=C hydrogen bonds involving proline residues in helices, *J. Phys. Chem. B 108*, 18065-18072.

(5) Ringer, A. L., Senenko, A., and Sherrill, C. D. 2007. Models of S/p interactions in protein structures: Comparison of the H_2S-benzene complex with PDB data, *Protein Science 16*, 2216-2223.

(6) Dougherty, D. A. 1996. Cation-pi interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp, *Science 271*, 163-168.

(7) Gallivan, J. P., and Dougherty, D. A. 1999. Cation-pi interactions in structural biology, *Proc. Natl. Acad. Sci. U. S. A. 96*, 9459-9464.

(8) Bartlett, G. J., Choudhary, A., Raines, R. T., and Woolfson, D. N. (2010) n-->pi* interactions in proteins, *Nat. Chem. Biol. 6*, 615-620.

(9) Sinnokrot, M. O., and Sherrill, C. D. 2006. High-accuracy quantum mechanical studies of pi-pi interactions in benzene dimers, *J. Phys. Chem. A 110*, 10656-10668.

(10) Jurecka, P., Sponer, J., Cerny, J., and Hobza, P. 2006. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs, *Phys. Chem. Chem. Phys. 8*, 1985-1993.

(11) Marsili, S., Chelli, R., Schettinoab, V., and Procacci, P. 2008. Thermodynamics of stacking interactions in proteins, *Phys. Chem. Chem. Phys. 10*, 2673–2685.

(12) Jackson, M. R., Beahm, R., Duvvuru, S., Narasimhan, C., Wu, J., Wang, H. N., Philip, V. M., Hinde, R. J., and Howell, E. E. 2007. A preference for edgewise interactions between aromatic rings and carboxylate anions: the biological relevance of anion-quadrupole interactions, *J. Phys, Chem. B 111*, 8242-8249.

(13) Quinonero, D., Garau, C., Rotger, C., Frontera, A., Ballester, P., Costa, A., and Deya, P. M. 2002. Anion-pi Interactions: do they exist?, *Angew. Chem. Int. Ed. Engl. 41*, 3389-3392.

(14) Garau, C., Quinonero, D., Frontera, A., Ballester, P., Costa, A., and Deya, P. M. 2003. Dual binding mode of s-triazine to anions and cations, *Org. Lett. 5*, 2227-2229.

(15) Garau, C., Frontera, A., Quinonero, D., Ballester, P., Costa, A., and Deya, P. M. (2003) A topological analysis of the electron density in anion-pi interactions, *Chemphyschem 4*, 1344-1348.

(16) Gallivan, J. P., and Dougherty, D. A. 1999. Can lone pairs bind to a pi system? The water···hexafluorobenzene interaction, *Org. Lett. 1*, 103-105.

(17) Alkorta, I., Rozas, I., and Elguero, J. 1997. An Attractive Interaction Between the Pi-Cloud of C6F6 and Electron-Donor Atoms, *J. Org. Chem. 62*, 4687-4691.

(18) Mascal, M., Armstrong, A., and Bartberger, M. D. 2002. Anion-aromatic bonding: a case for anion recognition by pi-acidic rings, *J. Am. Chem. Soc. 124*, 6274-6276.

(19) Kim, D., Tarakeshwar, P., and Kim, K. 2004. Theoretical Investigations of Anion-pi Interactions: The Role of Anions and the Nature of Pi Systems, *J. Phys. Chem. A 108*, 1250-1258.

(20) Danten, Y., Tassaing, T., and Besnard, M. 1999. On the Nature of the Water-Hexafluorobenzene Interaction, *J. Phys. Chem. A 103*, 3530-3534.

(21) Garau, C., Frontera, A., Quinonero, D., Ballester, P., Costa, A., and Deya, P. 2004. Cation-pi versus anion-pi interactions: Energetic, charge transfer and aromatic aspects, *J. Phys. Chem. A 108*, 9423-9427.

(22) Quinonero, D., Garau, C., Frontera, A., Ballester, P., Costa, A., and Deya, P. M. 2005. Structure and binding energy of anion-pi and cation-pi complexes; A comparison of MP2, RI-MP2, DFT and DF-DFT methods, *J. Phys. Chem. A 109*, 4632-4637.

(23) Garau, C., Frontera, A., Quinonero, D., Ballester, P., Costa, A., and Deya, P. 2004. Cation-pi vs anion-pi interactions: A complete pi-orbital analysis, *Chem. Phys. Letts 399*, 220-225.

(24) Garau, C., Frontera, A., Quinonero, D., Ballester, P., Costa, A., and Deya, P. M. 2004. Cation-pi versus anion-pi interactions: A comparative ab initio study based on energetic, electron charge density and aromatic features, *Chem. Phys. Letters 392*, 85-89.

(25) Garau, C., Frontera, A., Quinonero, D., Ballester, P., Costa, A., and Deya, P. M. 2004. Anion-pi interactions, *Recent Res. Devel. Chem. Physics 5*, 227-255.

(26) Rosokha, Y. S., Lindeman, S. V., Rosokha, S. V., and Kochi, J. K. 2004. Halide recognition through diagnostic "anion-pi" interactions: molecular complexes of Cl-, Br-, and I- with olefinic and aromatic pi receptors, *Angew. Chem. Int. Ed. Engl. 43*, 4650-4652.

(27) de Hoog, P., Gamez, P., Mutikainen, I., Turpeinen, U., and Reedijk, J. 2004. An aromatic anion receptor: Anion-pi interactions do exist, *Angew. Chem. Int. Ed. Engl. 43*, 5815-5817.

(28) Quinonero, D., Garau, C., Frontera, A., Ballester, P., Costa, A., and Deya, P. M. 2002. Counterintuitive interactions of anions with benzene derivatives, *Chem. Phys. Letters 359*, 486-492.

(29) Frontera, A., Saczewski, F., Gdaniec, M., Dziemidowicz-Borys, E., Kurland, A., Deya, P. M., Quinonero, D., and Garau, C. 2005. Anion-pi interactions in cyanuric acids: a combined crystallographic and computational study, *Chemistry 11*, 6560-6567.

(30) Berryman, O. B., Hof, F., Hynes, M. J., and Johnson, D. W. 2006. Anion-pi interaction augments halide binding in solution, *Chem. Commun. (Camb.)*, 506-508.

(31) Thomas, K. A., Smith, G. M., Thomas, T. B., and Feldmann, R. J. 1982. Electronic distributions within protein phenylalanine aromatic rings are reflected by the three-dimensional oxygen atom environments, *Proc. Natl. Acad. Sci. U. S. A. 79*, 4843-4847.

(32) Burley, S. K., and Petsko, G. A. 1988. Weakly polar interactions in proteins, *Adv. Protein Chem. 39*, 125-189.

(33) Duan, G., Smith, J., V.D., and Weaver, D. F. 2001. Characterization of aromatic-thiol pi-type hydrogen bonding and phenylalanine-cysteine side chain interactions through *ab initio* calculations and protein database analyses, *Mol. Phys. 99*, 1689-1699.

(34) Meyer, E. A., Castellano, R. K., and Diederich, F. 2003. Interactions with aromatic rings in chemical and biological recognition, *Angew. Chem. Int. Ed. Engl. 42*, 1210-1250.

(35) Singh, J., and Thornton, J. 1992. *Atlas of Protein Side-Chain Interactions*, Volumes 1 & 2, IRL Press at Oxford University Press, Oxford.

(36) Shi, Z., Olson, C. A., and Kallenbach, N. R. 2002. Cation-pi interaction in model alpha-helical peptides, *J. Am. Chem. Soc. 124*, 3284-3291.

(37) Olson, C. A., Shi, Z., and Kallenbach, N. R. 2001. Polar interactions with aromatic side chains in alpha-helical peptides: Ch...O H-bonding and cation-pi interactions, *J. Am. Chem. Soc. 123*, 6451-6452.

(38) Shi, Z., Olson, C. A., Bell, A. J., Jr., and Kallenbach, N. R. 2002. Non-classical helix-stabilizing interactions: C-H...O H-bonding between Phe and Glu side chains in alpha-helical peptides, *Biophys. Chem. 101-102*, 267-279.

(39) Joughin, B. A., Green, D. F., and Tidor, B. 2005. Action-at-a-distance interactions enhance protein binding affinity, *Protein Sci. 14*, 1363-1369.

(40) Soga, S., Shirai, H., Kobori, M., and Hirayama, N. 2007. Use of amino acid composition to predict ligand binding sites, *J. Chem. Inf. Model 47*, 400-406.

(41) Koide, S., and Sidhu, S. S. 2009. The importance of being tyrosine: lessons in molecular recognition from minimalist synthetic binding proteins, *Chem. Biol. 4*, 325-334.

(42) Fellouse, F. A., Barthelemy, P. A., Kelley, R. F., and Sidhu, S. S. 2006. Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code, *J. Mol. Biol. 357*, 100-114.

(43) Fellouse, F. A., Esake, K., Birtalan, S., Raptis, D., Cancasci, V. J., Koide, A., Jhurani, P., Vasser, M., Koide, S., and Sidhu, S. S. 2007. High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries, *J. Mol. Biol. 373*, 924-940.

(44) Gorteau, V., Bollot, G., Mareda, J., Perez-Velasco, A., and Matile, S. 2006. Rigid Oligonaphthalenediimide Rods as Transmembrane Anion-p Slides, *J. Am. Chem. Soc. 128*, 14788–14789.

(45) Ioanoviciu, A., Meharenna, Y. T., Poulos, T. L., and Ortiz de Montellano, P. R. 2009. DevS Oxy Complex Stability Identifies This Heme Protein as a Gas Sensor in *Mycobacterium tuberculosis* Dormancy, *Biochemistry 48*, 5839-5848.

(46) Juszczak, L. J., and Desamero, R. Z. B. 2009. Extension of the Tryptophan χ2,1 Dihedral Angle-W3 Band Frequency Relationship to a Full Rotation: Correlations and Caveats *Biochemistry 48*, 2777-2787.

(47) Dharmarajan, L., Case, C. L., Dunten, P., and Mukhopadhyay, B. 2008. Tyr235 of human cytosolic phosphoenolpyruvate carboxykinase influences catalysis through an anion-quadrupole interaction with phosphoenolpyruvate carboxylate *FEBS Journal 275*, 5810-5819.

(48) Sartorius, J., and Schneider, H. J. 1995. NMR-titrations with complexes between ds-DNA and indole derivatives including tryptophane containing peptides, *FEBS Lett. 374*, 387-392.

(49) Hobza, P. 2008. Stacking interactions, *Phys. Chem. Chem. Phys. 10*, 2581-2583.

(50) Hobohm, U., and Sander, C. 1994. Enlarged representative set of protein structures, *Protein Sci. 3*, 522-524.

(51) Hobohm, U., Scharf, M., Schneider, R., and Sander, C. 1992. Selection of representative protein data sets, *Protein Sci. 1*, 409-417.

(52) Schuettelkopf, A. W., and van Aalten, D. M. F. 2004. PRODRG - a tool for high-throughput crystallography of protein-ligand complexes, *Acta Crystallographica D60*, 1355-1363.

(53) Pace, C. N., Grimsley, G. R., and Scholtz, J. M. 2009. Protein ionizable groups: pK values and their contribution to protein stability and solubility, *J. Biol. Chem. 284*, 13285-13289.

(54) Shah, A. V., Walters, W. P., Shah, R., and Dolata, D. P. 1994. BABEL: A tool for converting between molecular structural data formats, in *Computerized Chemical Data Standards: Databases, Data Interchange, and Information Systems* pp 45-53, ASTM, Philadelphia.

(55) Schmidt, M., Baldridge, K., Boatz, J., Elbert, S., Gordon, M., Jensen, J., Koseki, S., Matsunaga, N., Nguyen, K., and al., e. 1993. General atomic and molecular electronic structure system, *J. Comput. Chem. 14*, 1347-1363.

(56) Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Zakrzewski, V. G., Montgomery, J., J. A. , Stratmann, R. E., Burant, J. C., Dapprich, S., Millam, J. M., Daniels, A. D., Kudin, K. N., Strain, M. C., Farkas, O., Tomasi, J., Barone, V., Cossi, M., Cammi, R., Mennucci, B., Pomelli, C., Adamo, C., Clifford, S., Ochterski, J., Petersson, G. A., Ayala, P. Y., Cui, Q., Morokuma, K., Malick, D. K., Rabuck, A. D., Raghavachari, K., Foresman, J. B., Cioslowski, J., Ortiz, J. V., Baboul, A. G., Stefanov, B. B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Gomperts, R., Martin, R. L., Fox, D. J., Keith, T., Al-Laham, M. A., Peng, C. Y., Nanayakkara, A., Gonzalez, C., Challacombe, M., Gill, P. M. W., Johnson, B., Chen, W., Wong, M. W., Andres, J. L., Gonzalez, C., Head-Gordon, M., Replogle, E. S., and Pople, J. A. 1998., Gaussian, Inc., Pittsburgh

(57) Kitaura, K., and Morokuma, K. 1976. New energy decomposition scheme for molecular interactions within Hartree-Fock approximation, *Intl. J. Quantum Chem. 10*, 325-340.

(58) Ishida, K., Morokuma, K., and Komornicki, A. 1977. The intrinsic reaction coordinate. An ab initio calculation for HNC→HCN and $H^-+CH_4→CH_4+H^-$, *J. Chem. Phys. 66* 2153-2156

(59) Aschi, M., Mazza, F., and Di Nola, A. 2002. Cation-pi interactions between ammonium ion and aromatic rings: An energy decomposition study, *J. Mol. Structures 587*, 177-188.

(60) Umeyama, H., and Morokuma, K. 1977. The origin of hydrogen bonding. An energy decomposition study *J. Am. Chem. Soc. 99*, 1316-1332.

(61) Dunning Jr., T. H. 1989. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. , *J. Chem. Phys. 90*, 1007-1023.

(62) Kendall, R. A., Dunning Jr., T. H., and Harrison, R. J. 1992. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. , *J. Chem. Phys. 96*, 6796-6806.

(63) Boys, S., and Bernardi, F. 1970. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors., *Mol. Phys. 19*, 553-566.

(64) MacKerell Jr., A. D., Bashford, D., Bellott, M., Dunbrack Jr., R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher III, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., and Karplus, M. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins, *J. Phys. Chem. B 102*, 3586-3616.

(65) Walters, P., Stahl, M. at Department of Chemistry, University of Arizona, Tucson, AZ, 1994, modified by Gosper, J. at BRUNEL University, UK. pp ftp://joplin.biosci.arizona.edu/pub/Babel/.

(66) Chou, P. Y., and Fasman, G. D. 1974. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins., *Biochemistry 13*, 211-222.

(67) Chou, P. Y., and Fasman, G. D. 1974. Prediction of protein conformation, *Biochemistry 13*, 222-245.

(68) Pokkuluri, P. R., Gu, M., Cai, X., Raffen, R., Stevens, F. J., and Schiffer, M. 2002. Factors contributing to decreased protein stability when aspartic acid residues are in b-sheet regions, *Protein Sci. 11*, 1687-1694.

(69) Niwa, T. S., and Ogino, A. 1997. Multiple regression analysis of the beta-sheet propensity of amino acids, *Journal of Molecular Structure (Theochem) 419*, 155-160.

(70) Smith, C. K., and Regan, L. 1997. Construction and design of b-sheets, *Acc. Chem. Res. 30*, 153-161.

(71) Richardson, J. S., and Richardson, D. C. 2002. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation, *Proc. Natl. Acad. Sci. U. S. A. 99*, 2754-2759.

(72) Pintar, A., Carugo, O., and Pongor, S. 2002. CX, an algorithm that identifies protruding atoms in proteins, *Bioinformatics 18*, 980-984.

(73) Mihel, J., Sikic, M., Tomic, S., Jeren, B., and Vlahovicek, K. 2008. PSAIA - protein structure and interaction analyzer, *BMC Struct. Biol. 8*, 21.

(74) Fletterick, R. J., Schroer, T., and Matela, R. J. 1985. *Molecular Structure, Macromolecules in 3D*, Blackwell Scientific Publications, Oxford.

(75) Xu, Y., Shen, J., Zhu, W., Luo, X., Chen, K., and Jiang, H. 2005. Influence of the Water Molecule on Cation-p Interaction: Ab Initio Second Order Møller-Plesset Perturbation Theory (MP2) Calculations, *J. Phys. Chem. B 109*, 5945-5949.

(76) Prajapati, R. S., Sirajuddin, M., Durani, V., Sreeramulu, S., and Varadarajan, R. 2006. Contribution of Cation-p Interactions to Protein Stability, *Biochemistry 45*, 15000-15010.

(77) Quinonero, D., Deya, P. M., Carranza, M. P., Rodriquex, A. M., Jalon, F. A., and Manzano, B. R. 2010. Experimental and computational study of the interplay between C–H/p and anion–p interactions†, *Dalton Trans. 39*, 794–806.

(78) Lucas, X., Estarellas, C., Escudero, D., Frontera, A., Quinonero, D., and Deya, P. M. 2009. Very long-range effects: Cooperativity between anion–p and hydrogen-bonding interactions, *ChemPhysChem 10*, 2256–2264.

(79) Quinonero, D., Frontera, A., Garau, C., Ballester, P., Costa, A., and Deya, P. M. 2006. Interplay between cation–p, anion–p and p–p interactions, *ChemPhysChem 7*, 2487–2491.

(80) Das, A., Choudhury, S. R., Dey, B., Yalamanchili, S. K., Helliwell, M., Gamez, P., Mukhopadhyay, S., Estarellas, C., and Frontera, A. 2010. Supramolecular assembly of Mg(II) complexes directed by associative lone pair-pi/pi-pi/pi-anion-pi/pi-lone pair interactions., *J. Phys. Chem. B. 114*, 4998-5009.

(81) Gilis, D., Biot, C., Buisine, E., Dehouck, Y., and Rooman, M. 2006. Development of novel statistical potentials describing cation-pi interactions in proteins and comparison with semiempirical and quantum chemistry approaches., *J. Chem. Inf. Model 46*, 884-893.

(82) Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., and Zardecki, C. 2002. The protein data bank, *Acta Crystallogr. D Biol. Crystallogr. 58*, 899-907.

(83) Atwood, J., and Steed, J. 1997. Structural and topological aspects of anion coordination, in *Supramolecular Chemistry of Anions* (Bianchi, A., Bowman-James, K., and Garcia-Espana, E., Eds.) pp 147-215, Wiley-VCH, New York.

(84) Levy, Y., and Onuchic, J. N. 2006. Water mediation in protein folding and molecular recognition, *Annu. Rev. Biophys. Biomol. Struct. 35*, 389-415.

(85) Chong, L. T., Dempster, S. E., Hendsch, Z. S., Lee, L. P., and Tidor, B. 1998. Computation of electrostatic complements to proteins: a case of charge stabilized binding, *Protein Sci. 7*, 206-210.

(86) Ponder, J. W., Wu, C., Ren, P., Pande, V. S., Chodera, J. D., Mobley, D. L., Schnieders, M. J., Haque, I., Lambrecht, D. S., DiStasio, J., R.A., Head-Gordon, M., Clark, G. N. I., Johnson, M. E., and Head-Gordon, T. 2010. Current status of the AMOEBA polarizable force field, *J. Phys. Chem. B 114*, 2549-2564.

(87) Lopes, P. E. M., Roux, B., and MacKerell Jr., A. D. 2009. Molecular modeling and dynamics studies with explicit inclusion of electronic polarizability. Theory and applications, *Theor. Chem. Acc. 124*, 11-28.

(88) Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., and Tramontano, A. 2009. Critical assessment of methods of protein structure prediction—Round VIII, *Proteins 77(Suppl 9)*, 1-4.

CHAPTER 3.2. STAAR: STATISTICAL ANALYSIS OF AROMATIC RINGS

A version of this chapter was originally published by David D. Jenkins, Jason B. Harris, Elizabeth E. Howell, Robert J. Hinde, and Jerome Baudry:

David D. Jenkins, Jason B. Harris, Elizabeth E. Howell, Robert J. Hinde, Jerome Baudry. "STAAR: STatistical Analysis of Aromatic Rings"**.** *Journal of Comp. Chemistry* (2013). 34(6):518-22

The work and writing in this chapter was contributed to by all authors. J. Baudry, E. Howell, and R.J. Hinde served as faculty advisors. J.B. Harris designed project, served as project manager, senior student advisor, and code designer. D. Jenkins served as the C++ programmer, generated data, and drafted the methods and results sections for the manuscript.

Abstract

The STAAR (Statistical Analysis of Aromatic Rings) program allows for an automated search for anion-π interactions between phenylalanine residues and carboxylic acid moieties of neighboring aspartic acid or glutamic acid residues in Protein Data Bank (PDB) structures. The program is written in C++ and is available both as a standalone code and through a web implementation that allows users to upload and analyze biomolecular structures in PDB format. The program outputs lists of Phe/Glu or Phe/Asp pairs involved in potential anion-π interactions, together with geometrical (distance and angle between the Phe's center of mass and Glu or Asp's center of charge) and energetic (quantum mechanical Kitaura-Morokuma interaction energy between the residues) descriptions of each anion-π interaction. Application of the program on the latest content of the PDB shows that anion-π interactions are present in thousands of protein structures and can possess strong energies, as low as -8.72 kcal/mol.

Introduction

Anion-π interactions, in which negatively charged species interact with the positively charged edge of resonant groups, represent a common but unrecognized and poorly characterized type of non-bonded interaction in chemistry that can exhibit quantum mechanical energies as strong as -23.5 kcal/mol.[1-7] There is a sustained interest in data-mining known databases of chemical structures to identify and characterize anion-π interactions. Several searches have found such interactions in the Cambridge Structural Database (CSD).[2–8] The potential importance of such interactions in biomolecules has stimulated searches in the Protein Data Bank (PDB) for pairs involving aromatic side chains and isolated ionic species (Cl^-, PO_4^{-3}, NO_3^-, Br^-, F^- and ClO_4^-).[9-10] Using a program specifically written for that purpose, it was found that potentially strong (as extrapolated from geometrical criteria) and functionally important [11] anion-π interactions are found in the PDB but only in small numbers in protein structures, largely because of the relatively low numbers of 'isolated ions' in the PDB[10]. Our previous work has, on the contrary, identified roughly 19,000 potential anion-π interactions involving residue/residue interactions in the PDB, between the neutral resonant ring of phenylalanine (Phe) and carboxylic acid moieties of aspartic acid (Asp) or glutamic acid (Glu) side chains.[12] The quantum mechanical interaction energies of

simplified models of these pairs (in which Phe was represented as benzene and Glu or Asp was represented as formate anion) were calculated to be as strong as -7.27 kcal/mol, suggesting that anion-π interactions in proteins can be common and relatively strong.[12] While their local protein environment and solvation may weaken these interaction energies, such common and strong side-chain/side-chain anion-π interactions in biomolecules may contribute both to the overall stability of biomolecular structures and complexes and to their function through substrate binding and protein-protein interactions. This makes the search for, and characterization of, anion-π interactions in biomolecules of great interest to the structural biology community. The present work describes the STAAR (Statistical Analysis of Aromatic Rings) program and our implementation of the program for the entire PDB, allowing a systematic search for Phe/Asp or Phe/Glu anion-π side chain interactions, followed by the quantum mechanical calculation of the corresponding benzene/formate interaction energies. The program is applied to the latest content of the PDB and the results demonstrate the efficiency of this computational program. The STAAR program is available free of charge from the website, http://staar.bio.utk.edu.

Methodology

Process

The flow chart of the STAAR program is given in Figure 3.2.1. This program will read PDB entries into memory using a specially developed C++ parser.[9,13] If the resolution of a given PDB structure is coarser than a specified threshold, set to 3Å in the case of the analysis reported here, the structure is not processed. Otherwise, the structure will be further processed. PDB entries that contain multiple models, such as NMR structures, are split up into different entries and analyzed individually. Within a PDB entry, all chains are treated individually, which allows inter- and intra-chain interactions to be analyzed. In each of the entries and each of the chains, the phenylalanine (Phe), and aspartic acid (Asp) or glutamic acid (Glu) amino acid pairs are identified as described below.

First, the PDB file is parsed for the each of the three residues, and they are stored into a vector of AminoAcid objects. Secondly, the potential anion-π pairs (Phe-Asp and Phe-Glu) are then identified as follows: for each of the pairs, the side chains of Phe and Asp/Glu are simplified to benzene and formate, respectively. The center of mass of a benzene moiety is calculated as the average of the coordinates of its atoms as shown in equation 1, where $\mathbf{c}_i$ is the coordinate vector for the i^{th} carbon.

$$\overrightarrow{CM} = \frac{1}{6} * \sum_{i \in Carbons} \overrightarrow{c_i} \qquad \textbf{(1)}$$

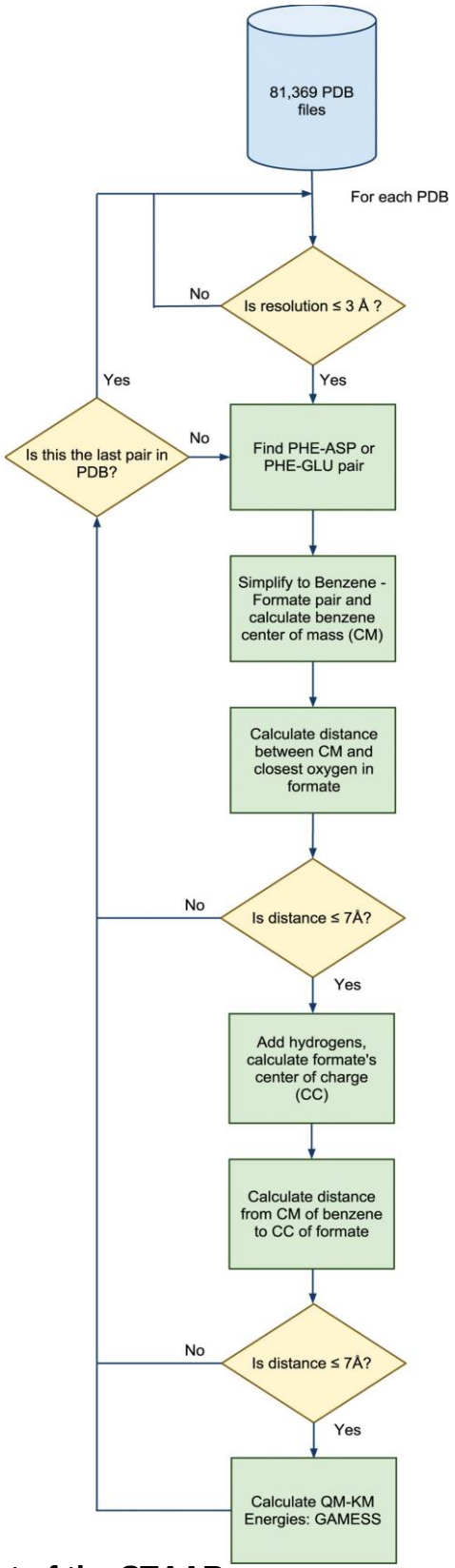


Figure 3.2.1. Flowchart of the STAAR program.

The program calculates the distance from the benzene center of mass to each of the oxygen atoms of the formate, and identifies the shorter of the two distances. If this distance is greater than a specified distance threshold, (7Å in the case of the analysis reported here), no anion-π interaction is defined and the analysis continues with the next potential pair. If this distance is less than or equal to this threshold, empty valences are filled with hydrogen atoms using the OpenBabel library.[14] The center of charge of the formate is then calculated using equation 2:

$$\overrightarrow{CC} = \overrightarrow{c_C} + \frac{\vec{v} * l}{\|\vec{v}\|}; \qquad \vec{v} = \overrightarrow{c_C} - \overrightarrow{c_H} \qquad \textbf{(2)}$$

where $\overrightarrow{c_C}$ and $\overrightarrow{c_H}$ are the x,y,z coordinates of the carbon and hydrogen atoms in the formate, $\overrightarrow{CC}$ is the center of charge, and l is the distance between the carbon atom and the formate center of charge, which was assigned a value of 0.632469 Å.[15] The distance and angle between the (benzene) center of mass and the (formate) center of charge are calculated from the x,y,z coordinates obtained by formula 1 and 2 above. After this point, STAAR outputs the GAMESS command files needed for quantum mechanical (QM) calculations (Table 3.2.1) and a comma separated file in csv format containing the amino acid information (residue names, locations, chain IDs, model number PDB ID) distances, and angles. The analysis is repeated with the next anion-π potential pair until the end of the PDB file and the process continues with the next structure.

Table 3.2.1. GAMESS input parameters.

```
$CONTRL SCFTYP = RHF RUNTYP = EDA ICHARG = -1 MULT =
  1 COORD = CART MAXIT = 200 $END
$SYSTEM TIMLIM = 1000 $END
$BASIS GBASIS = TZV $END
$GUESS GUESS = HUCKEL $END
$SCF SOSCF = .F. DAMP = .T. SHIFT = .T. DEM = .F. $END
$MOROKM IATM(1) = 12,4 ICHM(1) = 0,-1 $END
```

After all of the anion-π pairs are identified, quantum mechanical (QM) calculations are performed by GAMESS[16,17] to calculate the Kitaura-Morokuma[18] energy decomposition (KM) of each pair using the triple-zeta TZV basis set supplied with the GAMESS program. This is done using an embarrassingly parallel approach where each job performs the KM calculations for a single pair. The QM results are parsed out and combined with the STAAR csv file for analysis. The KM energy decomposition has been shown to highly correlate with HF and MP2 *ab initio* calculations of the interaction energies (slope= 0.92, χ^2 = 0.988 for correlation with HF calculations)[12]. At the same

time, KM energy decomposition calculations are much faster than *ab initio* HF or MP2 calculations and hence can be used to process the large number of pairs identified here.

Web Service

The process described above has been performed on the entire PDB database (81,369 PDB entries as of May 9, 2012) and the results are stored in a web-based, freely available and searchable database: http://staar.bio.utk.edu/. The user can search for the anion-π pairs identified for a specified structure by searching the database using its PDB code. The results are viewable in tables separated by model (if applicable) and ordered by ascending interaction energies. Each PDB page consists of tables containing the residue pair information (location, chain ID, distance, angle, and total energy) and links to export the results to a csv or tsv (tab separated file). The site includes the ability for a user to write their own script to copy PDB results to their own system in the form of a csv or tsv file. An example script is provided on http://staar.bio.utk.edu/search.php. Results can be browsed and sorted by PDB ID, resolution, number of pairs, and minimum energy and filtered by structure resolution.

Users can additionally submit their own PDB files to the STARR program (Figure 3.2.2). The user input is limited to uploading the structure and downloading the results. The search page contains a form requesting an email address and a path to a PDB file that can be uploaded using either an ASCII or a gzipped (.gz) version. When the job is complete, the webserver will email the user with a link to download results. Results are usually available within 10-30 minutes of submission. A Perl script on the search page is provided for the user to submit a file to the server without the web interface.

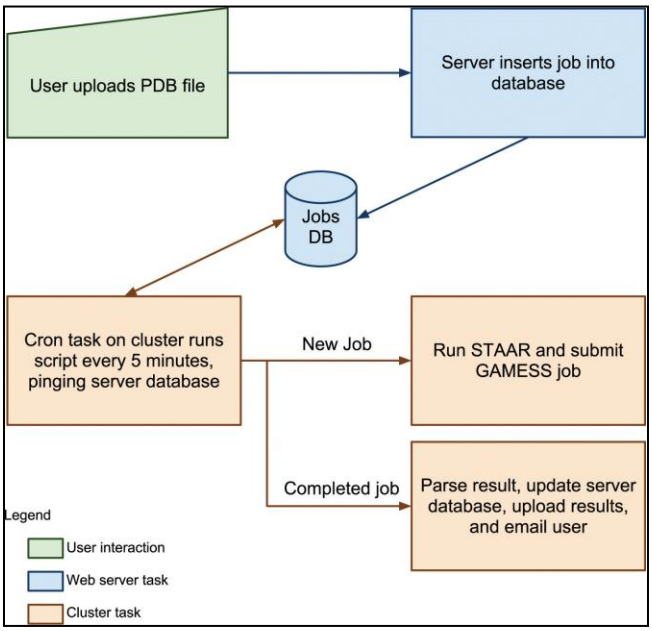


Figure 3.2.2. Web implementation of STAAR.

Code

The code is entirely written in C++ and is freely downloadable as open source. It is available for download, compilation, and installation on a Unix machine. The code has been successfully tested using the GNU GCC g++ compiler version 4.1.2 on various Linux versions and Intel's icpc compiler version 11.1 on Red Hat Enterprise Linux. To perform KM energy calculations, OpenBabel and GAMESS (downloaded separately) will need to be installed on the user's side.[14,16,17] Installation instructions and an example of how to run the STAAR program are available on the website.

Results

The code was run on the May 9, 2012 version of the Protein Data Bank, which consists of 81,369 entries. Running through the STAAR C++ code to identify potential anion-π pairs in all of the PDBs, restricting the search to resolutions of 3Å or finer, took approximately 1 minute by submitting 200 single core jobs to the Newton High Performance Computing system (https://newton.utk.edu/) to parse through the files in parallel. STAAR identified 818,066 pairs with a distance between the benzene center of mass and the center of charge in formate within a threshold of 7Å that were then processed by GAMESS to perform the KM energy decomposition. Each GAMESS run takes around 1 minute to complete, resulting in a complete processing of all identified anion-π pairs in 2.5 days. This was achieved by running the QM calculations in parallel, submitting single core jobs to the cluster, each job containing 10 GAMESS runs, and corresponds to a 227 fold speedup over serial calculations. Out of these 818,066 potential AQ pairs identified and processed by GAMESS, a total of 637 calculations did not converge. In addition, values computed for 42,800 pairs were not included as those calculations gave in a high order coupling mix energy with a magnitude greater than 0.25 kcal/mol, indicating a QM result that cannot be readily interpreted in terms of the KM energy decomposition. An additional 181 GAMESS runs did not yield results due to OpenBabel adding improperly hydrogen atoms to the molecules. This leads to usable data for 774,448 pairs in 65,452 PDB entries, representing 226,046 protein structure 'models' (one PDB ID can have several protein chains due to oligomerization and/or several copies of the same protein chain). This yields an average of ~3.4 anion-π pairs per protein structure, that fulfill the geometrical and calculation convergence criteria and are listed on the website. The entire PDB database contains over 8,000 NMR protein structures, each containing several models, and entries that contain multiple chains and copies of the same protein chain. Processing only the 4,272 non-redundant protein structures obtained by X-ray crystallography in the PDBselect database leads to 10,390 pairs in 4,277 protein structures, i.e. 2.43 pairs per protein.[19] The benzene-formate angles were distributed between all angles (0° to 90° range), with a preference for lower angle values. The shortest distance was 3.22 Å, as shown in Figure 3.2.3. The distribution of distances shows a nearly linear increase in the number of pairs as a function of distance. Figure 3.2.4 shows the number and distribution of KM energies for these pairs. The calculated energies range between -8.72 and +4.17 kcal/mol, with a most populated bin at -1.8 kcal/mol. About 81% of the anion-π pairs are calculated to have negative energies, i.e., are *a priori* stabilizing the corresponding protein structures.

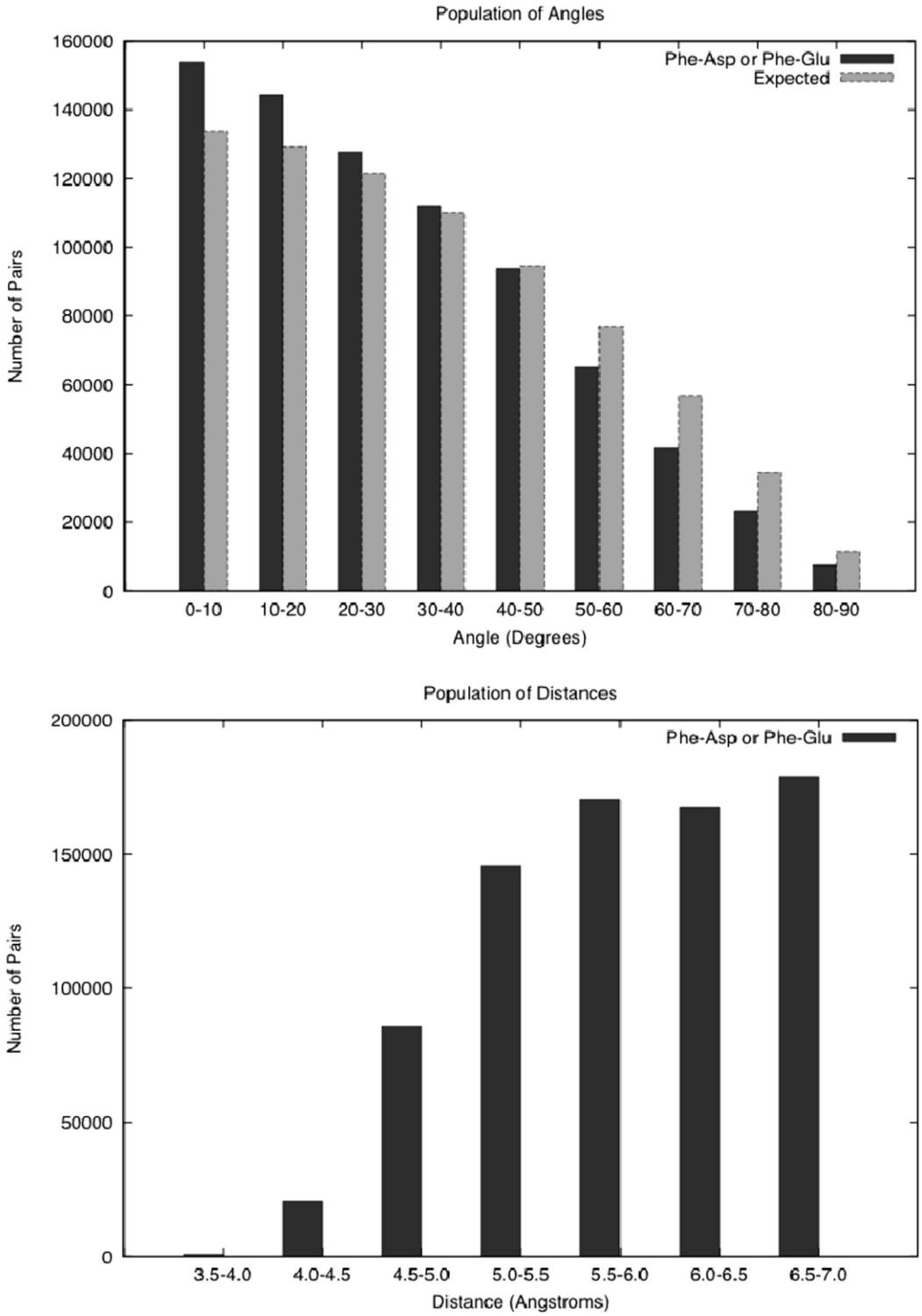


Figure 3.2.3. Angle and distance distributions for STAAR-identified anion-p pairs in the PDB. The gray bars for the angle plot describe the expected number of pairs based on statistical considerations. Here, the number of pairs is predicted based on two times the spherical sector volumes swept out by the designated angle divided by the total spherical volume.

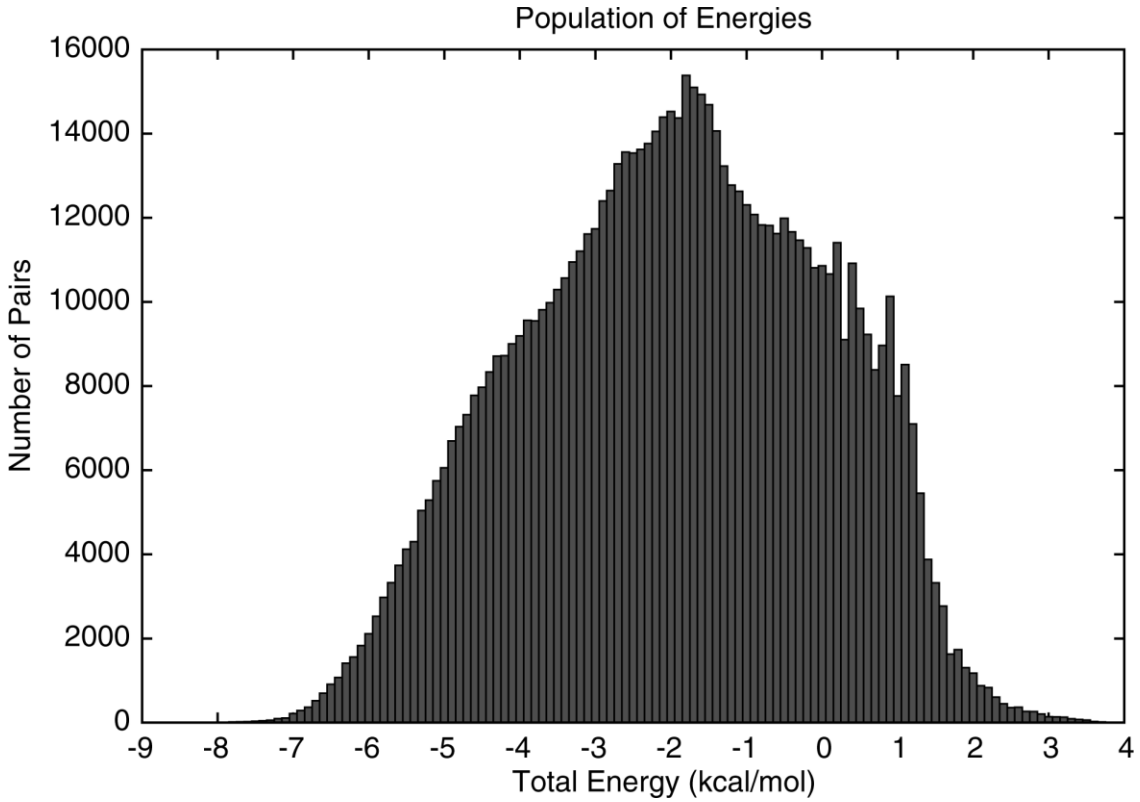


Figure 3.2.4. Distribution of KM energies for STAAR-identified anion-π pairs in the PDB.

Conclusion

The STAAR program can identify anion-π interactions in a large structural database of biomolecules. The program is freely available for download through our web interface, http://staar.bio.utk.edu. Applying the program on the most recent content of the PDB demonstrates the high prevalence and relatively strong anion-π energies involving side-chain/side-chain interactions in biomolecules. Future development of the program will involve extension of the code to include tryptophan and tyrosine residues. These aromatic groups will be used to calculate pairwise interaction energies with Asp and Glu as anions as well as the possible contribution of charge-dipole effects. Other future projects will investigate anion-π interaction in protein/ligand complexes, and the importance of hydration in the energetics of anion-π interactions.

Acknowledgment

The authors thank SCALE-IT (NSF grant 0801540) for funding this research. JB acknowledges support from a startup grant from the University of Tennessee/BCMB Department.

References

(1) Shi, G., Ding, Y., Fang, H. 2012. Unexpectedly strong anion–π interactions on the graphene flakes. *J. Comput. Chem*. 33, 1328-1337.

(2) Rosokha, Y. S., Lindeman, S. V., Rosokha, S. V., Kochi, J. K. 2004. Halide recognition through 'anion-π' interactions: Molecular complexes of Cl-,Br-, and I- with olefinic and aromatic π receptors. *Angew. Chem., Int. Ed. Engl.* 43, 4650-4652.

(3) de Hoog, P., Gamez, P., Mutikainen, I., Turpeinen, U., and Reedijk, J. 2004. An aromatic anion receptor: Anion-pi interactions do exist, *Angew. Chem. Int. Ed. Engl. 43*, 5815-5817.

(4) Quiñonero, D., Garau, C., Frontera, A., Ballester, P., Costa, A., Deyà, P. M. 2002. Counterintuitive interaction of anions with benzene derivatives. *Chem. Phys*. 359, 486-492.

(5) Frontera, A., Saczewski, F., Gdaniec, M., Dziemidowicz-Borys, E., Kurland, A., Deya, P. M., Quinonero, D., and Garau, C. 2005. Anion-pi interactions in cyanuric acids: a combined crystallographic and computational study, *Chemistry 11*, 6560-6567.

(6) Berryman, O. B., Hof, F., Hynes, M. J., and Johnson, D. W. 2006. Anion-pi interaction augments halide binding in solution, *Chem. Commun. (Camb.)*, 506-508.

(7) Frontera, A., Gamez, P., Mascal, M., Mooibroek, TJ., Reedijk, J. 2011. Putting anion-π interactions into perspective. *Angew. Chem. Int. Ed. Engl.* 50, 9564-9583.

(8) Hay, B. P., Custelcean, R. 2009. Anion-π Interactions in Crystal Structures: Commonplace or Extraordinary? *Cryst. Growth Des.* 2009, 9, 2539-2545.

(9) RCSB PDB, http://www.pdb.org (accessed July 2012).

(10) Robertazzi A., Krull, F., Knapp, E. W., Gamez, P. 2011. Recent Advances in Anion-pi Interactions. *Cryst. Eng. Comm.* 13, 3293-3300.

(11) Estarellas, C., Frontera, A., Quiñonero, D., Deyà, P.M. 2011. Relevant anion-π interactions in biological systems: the case of urate oxidase. *Angew. Chem. Int. Ed. Engl.* 50, 415-418.

(12) Philip, V., Harris, J., Adams, R., Nguyen, D., Spiers, J., Baudry, J., Howell, E.E., Hinde, R.J. 2011. A survey of aspartate-phenylalanine and glutamate-phenylalanine interactions in the protein data bank: searching for anion-π pairs. *Biochemistry* 50, 2939-2950.

(13) Berman, H., Henrik, K., Nakamura, H. 2003. Announcing the worldwide Protein Data Bank. *Nat. Struct. Mol. Biol.* 10, 980.

(14) O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R. 2011. Open Babel: an open chemical toolbox. *J Cheminform*. 6, 1–14.

(15) Jackson, M. R., Beahm, R., Duvvuru, S., Narasimhan, C., Wu, J., Wang, H. N., Philip, V. M., Hinde, R. J., and Howell, E. E. 2007. A preference for edgewise interactions between aromatic rings and carboxylate anions: the biological relevance of anion-quadrupole interactions, *J. Phys, Chem. B 111*, 8242-8249.

(16) Schmidt, M., Baldridge, K., Boatz, J., Elbert, S., Gordon, M., Jensen, J., Koseki, S., Matsunaga, N., Nguyen, K., and al., e. 1993. General atomic and molecular electronic structure system, *J. Comput. Chem. 14*, 1347-1363.

(17) Gordon, M. S., Schmidt, M. W. 2005. Advances in electronic structure theory: GAMESS a decade later. In *Theory and Applications of Computational Chemistry: The First Forty Years*, Elsevier Science, Amsterdam 1167-1189.

(18) Kitaura, K., and Morokuma, K. 1976. New energy decomposition scheme for molecular interactions within Hartree-Fock approximation, *Intl. J. Quantum Chem. 10*, 325-340.

(19) Griep, S., Hobohm, U. 2010. PDBselect 1992–2009 and PDBfilter-select. *Nucleic Acids Res.* 38, D318-D319.

CHAPTER 3.3. ANION-PI GEOMETRIES BETWEEN PROTEIN AND LIGAND STRUCTURES

Chapter 3 (Part III) represents unpublished work by Jason B. Harris, Aaron Mishtal, Caroline S. Rempe, Elizabeth E. Howell, Robert J. Hinde, and Jerome Baudry:

The work in chapter 3 (part III) was contributed to by all members. J. Baudry, E. Howell, and R.J. Hinde served as faculty advisors. J.B. Harris secured and managed project funds, recruited students and managed research plan, designed code, and generated data. Aaron Mishtal served as C++ programmer and Caroline S. Rempe assisted in code design and data analysis.

Abstract

Anion-pi interactions are emerging as an important noncovalent interaction in the stability and function of biological structures. Our previous work in Jackson et al., 2007, Philip et al., 2011 and Jenkins et al., 2013 established that energetically favorable anion-pi interactions occur between PHE and GLU/ASP sidechains for most structures represented in the Protein Data Bank (PDB) and also that favorable energies are associated for anions at nearly co-planar angles to aromatic rings. A new study of the PDB is now presented which establishes that approximately 74% of all co-crystallized ligands with a benzene-like (6-carbon) aromatic ring appear to be in energetically favorable anion-pi geometries relative to nearby GLU/ASP protein sidechains. These recent results along with our previous findings suggest that anion-pi interactions have a stabilizing role in both protein structure and ligand binding.

Introduction

Anion-pi, also referred to as anion-quadrupole, interactions are theorized to form between aromatic groups and anions. Aromatic functional groups are planar and ringed structures with delocalized pi electrons. The delocalization of electrons in an aromatic molecule create an electron density above and below the plane of the ring which can be described as a quadrupole (i.e., two opposite dipoles). The direction of the quadrupole creates a positive potential along the plane of an aromatic ring and this allows for favorable interactions with nearly co-planar anions. The basic principles of an anion-pi interaction are shown in Figure 3.3.1 between a benzene and a formate molecule.

Anion-pi interactions have been thought to exist for many years[1], and recently there has been a great interest in determining their role within biological structures[2,3]. This interest is largely due to the success of understanding the biological role of a closely related interaction called cation-pi. Cation-pi interactions occur between the negative pi clouds of aromatic rings and nearby cations, and they can stabilize protein structures by ~3 ± 1.5 kcal/mol[4,5]. The anion-pi interaction has been calculated to have energies as favorable as -8.72 kcal/mol[6].

The sidechains of protein amino acids contain polar, hydrophobic, aromatic, cationic, or anionic functional groups which contribute to protein structure and function. Stabilizing interactions between these groups in protein structure can be studied statistically from the many experimental structures existing in the Protein Data Bank (PDB). The STAAR (**ST**atistical **A**nalysis of **A**romatic **R**ings) program has been used in previous iterations of our work to look at the distribution of aromatic phenylalanine and

anionic glutamate/aspartate amino acid sidechains in both the entire PDB[6]and a non-redundant subset of the PDB[7]. Our previous work also suspected a preference of edge-wise interactions[8], with respect to the plane of aromatic rings, which was confirmed in Philip et al., 2011 and Jenkins et al., 2013. Those works also identified a substantial presence for these interactions in protein structures, and these are listed at the STAAR website (http://staar.bio.utk.edu).

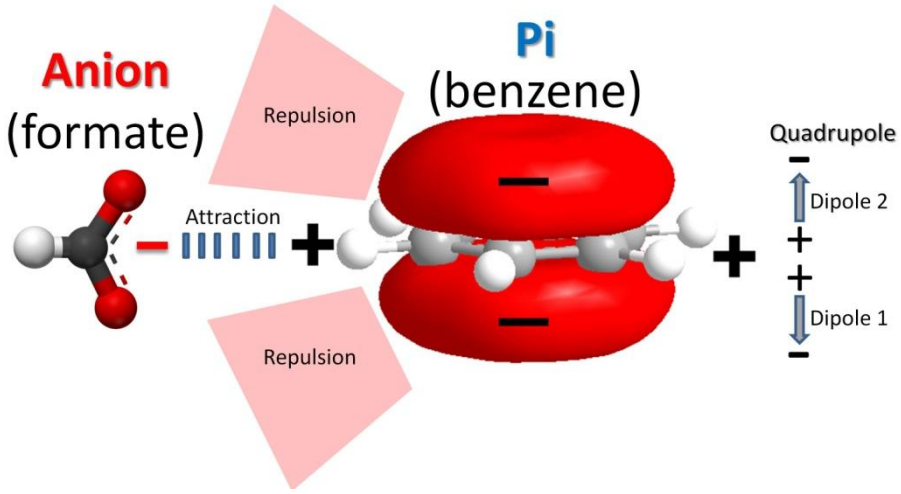


Figure 3.3.1. Principles of Anion-pi Interaction between benzene and formate.

Other studies that follow our work have lead to characterizing new roles for anion-pi interactions. Evidence for anion-pi interactions in nucleic acid structures was presented in Chakravartya et al., 2012[9]. The use of an anion-pi interaction to stabilize the position of amino acids for a catalytic site was presented by the Herschlag group in Schwans et al., 2013[10]. Robertazzi et al., 2011[11] presented work to investigate anion-pi interactions between free negatively charged ions and aromatic amino acids. There is still much work left at uncovering the role and significance of anion-pi interactions.

Noncovalent interactions are understood to be fundamental in ligand binding, and the significance of anion-pi interactions in protein structure has lead to studies aimed at understanding the role of this interaction in protein-ligand structure. The 2011 anion-pi study by Robertazzi with free ionic species (Cl^-, PO_4^{-3}, NO_3^-, Br^-, F^- and ClO_4^-) determined that a low number of negative ions were distributed near aromatic amino acid sidechains; however, the conclusions were weak due to not finding many free ionic species in the PDB. In order to provide a more robust study of anion-pi interactions between protein and ligands, the STAAR code has now been expanded to search for six-carbon aromatic groups on ligand structures. This allowed searching for anion-pi interactions between anionic formate groups from amino acid sidechains (GLU/ASP) and benzene-like aromatic groups (6-carbon rings) from ligand molecules. We now

present data to show that a large number of potentially favorable anion-pi interactions exist between protein-ligand structures of this type in the PDB.

Methods

General Methodology

The previous STAAR program only supported searching for anion-pi interactions between the functional groups of amino acid sidechains. PDB formatting guidelines for amino acids allowed for easy identification of protein sidechain atoms in the previous STAAR code; however, no formatting exists for easily identifying functional groups of interest in ligand structures. To overcome this issue, a set of geometrical rules were developed that allow STAAR to find benzene-like aromatic groups in ligand structures. This was coded in C++ and inserted as a search module in the existing STAAR program. The new STAAR program was used to search for GLU/ASP sidechains interacting with benzene-like aromatic rings in ligand structures, and this search was carried out on structures from the April 15th, 2013 version of the PDB.

Algorithm for Identifying 6-Carbon Aromatic Rings in Ligand Structures

Geometric rules were used to parse the heteroatoms (HET) atoms in the PDB and identify aromatic carbons in six-membered rings. Distances between a HET residues' carbon atoms were calculated and flagged when found within an approximate distance of 2.8 Å from each other, corresponding to the distance of opposite carbons in an benzene ring. The midpoints for each set of flagged carbon-carbon (C-C) pairs were then compared and C-C pairs having the same midpoint were defined as members a unique aromatic ring.

STAAR Program

Methods describing the overall STAAR program can be found in Philip et al. 2011 and Jenkins et al., 2013. Briefly, anion-pi pairs are parsed from PDB structure files using a distance cutoff of 7 Å. Anion-pi pairs are defined as benzene-like functional groups from protein (PHE) and ligand molecules and formate-like groups from protein molecules (ASP/GLU). Angles are calculated for each pair relative to the plane of the aromatic ring. Distances are calculated from the center of mass for the aromatic ring to the center of charge for the formate molecule.

Results

A search for anion-pi interactions in the PDB has been performed using an updated version of the STAAR program. This search has found 18,015 potential anion-pi interactions between protein-ligand complexes, defined by GLU/ASP amino acid sidehchains near 6-carbon aromatic rings on ligands. These results are significant since only 18,330 ligands were found to contain an aromatic ring, and also there were only 25,617 total aromatic rings. The average number of pairs found per aromatic ring was 0.70, and the average number of pairs found per ligand with at least 1 aromatic ring was 0.98. This statistical information is summarized in Table 3.3.1.

Table 3.3.1. Summary of Anion-pi Statistics.

Field	Description	Statistic
A.	PDB Structures with an Aromatic Ligand	6808
B.	Total interactions	18015
C.	Total Ligands with an Aromatic Ring	18330
D.	Total Aromatic Rings	25617
E.	Rate of Interactions/Aromatic Ring	0.70
F.	Rate of Interactions/Ligand with Aromatic Ring	0.98
G.	Total Favorable Interactions (Angle < 40°)	13477 or 75% of Total
I.	Rate of Favorable Interactions/Aromatic Ring	0.52 (75% of field E)
H.	Rate of Favorable Interactions/Ligand with Aromatic ring	0.74 (75% of field F)

Figure 3.3.1 shows the distribution of total interactions for ligands with a varying number of aromatic rings. For ligands with just a single aromatic ring, there was approximately a 9:12 (0.78) ratio of pairs to rings. The figure shows that this ratio becomes inverted for ligands with more than 1 aromatic ring whereby there is a higher rate of potential anion-pi interactions when ligands have multiple rings. This may suggest that proteins which bind compounds with multiple aromatic rings are adapted to take advantage of anion-pi interactions more frequently.

Our previous results in Philip et al., 2011 and Jenkins et al., 2013 provided a robust number of geometries for benzene-formate anion-pi pairs and their calculated QM energies. In those results, favorable energies were found for all pairs with an angle of less than 40 degrees. The distribution of angles for the current results are shown in Figure 3.3.3. Considering the number of anion-pi interactions that fall within a geometry of 40 degrees, we conclude that 13,477 favorable anion-pi pairs may exist out of the 18,015 total interactions evaluated (75% of Total). Based on 75% of geometries being considered favorable, it can also be estimated that 74% of all ligands, containing at least 1 aromatic ring, are also in favorable geometries (adjusting 0.98 by 75%). Likewise, ~52% of benzene-like rings (on ligands) are potentially involved in an anion-quadrupole interaction with GLU/ASP sidechains (on proteins). This information is summarized in Table 3.3.1.

Figure 3.3.4. shows the distribution of angles and distances for the 18,015 anion-pi pairs. In combination with Figure 3.3.3. it can be determined that a higher frequency of pairs are found for anion-pi pairs at low angles. This is expected to be the case if energies are more favorable between benzene-formate groups at low angles. Figure 3.3.4 also shows the average angles and distances for each binned group of pairs in 10 degree increments. Average pair distances vary for each bin, and favorable geometries (< 40 degrees) have greater average distances than less favorable geometries (> 40 degrees). The two outlier distances (< 1 Å) are the result of improperly formatted PDB entries. Most pairs fall within the range of 4 Å to 7 Å.

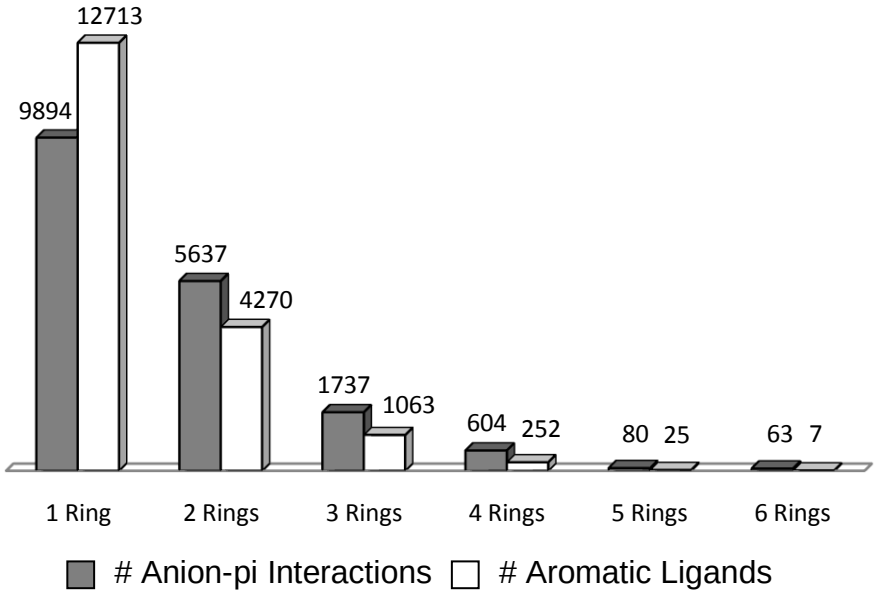


Figure 3.3.2. Anion-pi Pairs Involving Ligands with 1 to 6 Rings.

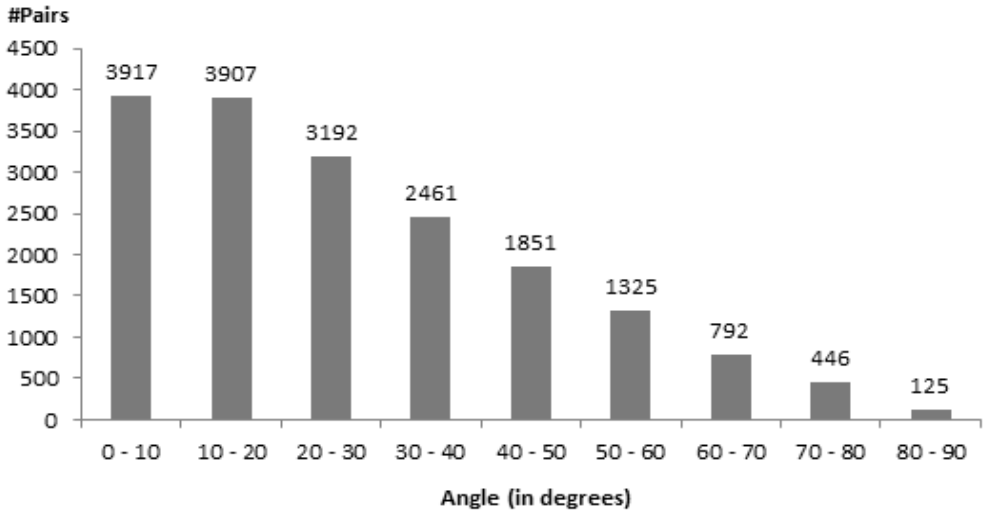


Figure 3.3.3. Anion-pi Pair Frequencies by Angle (10° increments).

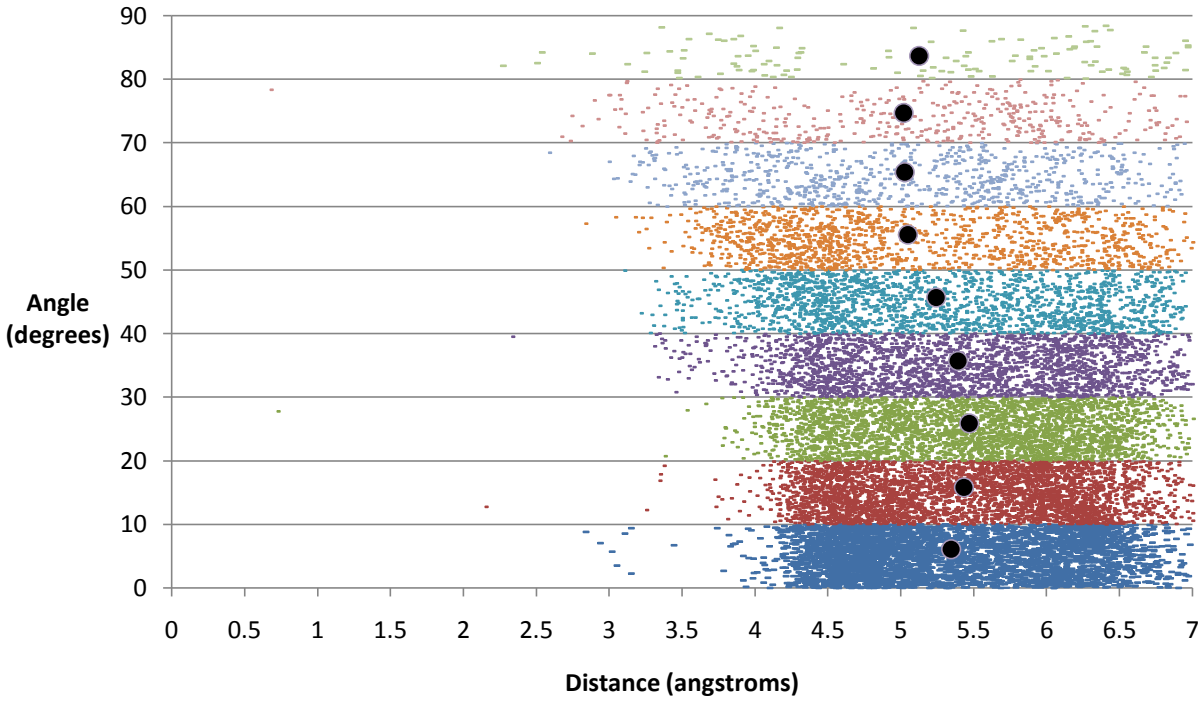


Figure 3.3.4. Anion-pi Distance and Angle Distribution. Pairs found in each angle grouping (binned in 10° increments) are colored the same. • A black circle represents the average distance and angle for each bin.

Discussion

The current analysis of the PDB included an upgrade to the STAAR program that allowed searching for benzene-like aromatic rings in ligand molecules which interact with formate-like groups in GLU/ASP sidechains. This analysis extends the description of potential anion-pi interactions to protein-ligand structures, and the results show a significant number of potentially favorable geometries for the two anion-pi functional groups (benzene and formate) currently under investigation. The results also confirm a preference for low angles between anion-pi pairs.

It is important to point out that the aromatic rings on ligand structure are often substituted with other functional groups which may contribute to some of the observed geometries investigated in this study. A post-processing of the 18,015 total interactions revealed that 1333 do not have any nearby charged groups within 2 bond distances of the aromatic ring. These simpler systems are not discussed in the scope of this study, but this may be useful to keep in mind for future studies that evaluate the reproducibility (e.g., molecular dynamics and docking) or experimental stability (e.g., mutagenesis) of these geometries.

Recent results from searching the PDB for anion-pi interactions suggest an important role for benzene-formate pairs in stabilizing biological structures. The current protein-ligand description of anion-pi pairs represents an important new functional role for this motif. This leads to further questions about the role of this interaction in modeling protein-ligand binding, and its implications in the fields of drug discovery and toxicology. Many drug-like compounds contain aromatic rings and bind to protein structures. It is expected that these questions will be evaluated in finer detail with case studies from interactions identified from this recent work. We plan to publish the updated STAAR code and a searchable list of all potential anion-pi pairs at the STAAR website (http:staar.bio.utk.edu).

Acknowledgment

The authors thank SCALE-IT (NSF grant 0801540) for funding this research.

References

(1) Thomas, K.A., Smith, G.M., Thomas, T.B., Feldmann, R.J. 1982 Electronic distributions within protein phenylalanine aromatic rings are reflected by the three-dimensional oxygen atom environments. *Proc Natl Acad Sci USA* 79, 4843–4847.

(2) Frontera, A., Gamez, P., Mascal, M., Mooibroek, T.J., Reedijk, J. 2011. Putting anion-π interactions into perspective. *Angew Chem Int Ed Engl* 50, 9564–9583.

(3) Schottel B.L., Chifotides H.T., Dunbar ,K.R. (2008) Anion-pi interactions. *Chem Soc Rev* 37, 68–83.

(4) Dougherty, D. A. 1996. Cation-pi interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp, *Science 271*, 163-168.

(5) Gallivan, J. P., and Dougherty, D. A. 1999. Cation-pi interactions in structural biology, *Proc. Natl. Acad. Sci. U. S. A. 96*, 9459-9464.

(6) Jenkins D.D., Harris, J.B., Howell, E.E., Hinde, R.J., Baudry, J. 2013. STAAR: STatistical Analysis of Aromatic Rings. *Journal of Comp. Chemistry*. 34, 518-22.

(7) Philip, V., Harris, J., Adams, R., Nguyen, D., Spiers, J., Baudry, J., Howell, E.E., Hinde, R.J. 2011. A Survey of Aspartate- Phenylalanine and Glutamate-Phenylalanine Interactions in the Protein Data Bank: Searching for Anion-π Pairs. *Biochemistry* 50, 2939-50.

(8) Jackson, M. R., Beahm, R., Duvvuru, S., Narasimhan, C., Wu, J., Wang, H. N., Philip, V. M., Hinde, R. J., and Howell, E. E. 2007. A preference for edgewise interactions between aromatic rings and carboxylate anions: The biological relevance of anion-quadrupole interactions. *J. Phys. Chem. B* 111, 8242–8249.

(9) Chakravartya, S., Shenga, Z., Iversona, B., Mooreb, B. 2012. η^{6}”-Type anion–π in biomolecular recognition. *FEBS Letters*. 586, 4180–4185.

(10) Schwans, J. P., Sunden, F., Lassila, J. K., Gonzalez, A., Tsai, Y., & Herschlag, D. 2013. Use of anion–aromatic interactions to position the general base in the ketosteroid isomerase active site. *Proceedings of the National Academy of Sciences*, *110*, 11308-11313.

(11) Robertazzi, A., Krull, F., Knapp, E. W., Gamez, P. 2011. Recent Advances in Anion-pi Interactions. *Cryst. Eng. Comm.* 13, 3293-3300.

CONCLUSION

All biological functions can in theory be understood by molecular interactions within and between chemical and biological molecules. Proteins and ligands are especially important classes of molecules that are important to structurally study in order to understand disease and toxicity at the molecular level. The work in this dissertation presented specific biological applications in drug discovery and toxicology where new and meaningful biological information about specific protein-ligand interactions was gained. Throughout this work, integrated methods were used to deal with several current challenges in molecular modeling that need to be resolved in order to increase the overall number and complexity of structures that can be understood at a systems-level by their molecular interactions.

In Chapter 1, a traditional small molecule discovery application for a single protein target was presented. The disease state under investigation was alpha-1-antitrypsin deficiency. Creating a structure-based predictive model of this system was challenging due to the desired structure being a theoretical intermediate state. This challenge was overcome through the integration of homology modeling, singe-protein docking, and an *in vitro* binding assay which allowed for a model of the intermediate state to be generated and then validated on a set of 80 compounds. The model was shown to have predictive abilities for rationalizing the binding of a newly found lead compound, and a larger virtual screen was then conducted to prioritize the future experimental testing of 16 putative compounds able to treat alpha-1-antrypsin deficiency.

Additional work related to Chapter 1 should include re-docking all validated compounds to variations of the intermediate state model. These conformations can be generated through either homology modeling or molecular dynamics, and they may offer better correlation with experiments as new screening results become available. Mutagenesis studies are also needed in order to further rationalize the predicted site(s) for polymerization inhibitors to bind. This information can be used to improve the accuracy of the predictive screening model and lead to its larger-scaled use. Also, sampling intermediate state structures through hybrid homology modeling of stable state structures should also be further evaluated since it may be a useful technique in expanding the number of biological structures that can be predictively modeled through molecular modeling.

In Chapter 2, a computational toxicology application was presented that modeled the multi-protein binding and metabolism of a small molecule. It was shown that the fate of a compound, PCB-30, could be predicted in a more cellular context by first modeling the compounds structural changes as it interacts with metabolic P450 enzymes and

then how modeling these structural changes effects its interactions with another protein, the human estrogen receptor. The presented work used multi-protein docking and *in vitro* experiments to model and validate the specific P450 metabolism and estrogenic activity of PCB-30's metabolites.

The work in Chapter 2 serves as a proof-of-concept and example for using molecular docking in future multi-protein and metabolism modeling efforts. Next-step studies should include expanding the number and diversity of ligands to be used as prototype molecules in this docking scheme. Additional protein targets such as other nuclear hormone receptors and cytochrome P450 enzymes should also eventually be considered.

In Chapter 3, biological roles and physical characteristics for an unconventional molecular interaction, anion-pi, were computationally examined. This involved data mining experimental protein structures deposited in the PDB for molecular groups associated with this interaction. This work uncovered significant biological roles for anion-pi interactions which include stabilizing protein and protein-ligand structures when found in edge-wise geometries.

These new functional descriptions of anion-pi interactions are expected to have an impact on improving the energetic calculations for future molecular modeling studies of proteins and ligands. Future work in this areas should involve exploring the contribution of this interaction on the stability and function in other classes of biological molecules. Experimental work should include mutagenesis studies, crystallization, and equilibrium unfolding analyses to understand the net energetic and structural effects of anion-pi interactions on molecular structure.

VITA

Jason Bret Harris was born 1985 in Stanton, VA. He graduated from Bristol Tennessee High School in 2004. He attended the University of Tennessee in Knoxville where he graduated in 2009 with a Bachelor of Arts degree in Biology (biochemistry, cellular, and molecular). He matriculated as an National Science Foundation fellow into the Graduate School of Genome Science and Technology at the University of Tennessee where he was advised by Jerome Baudry. During his graduate studies, he joined the Center for Molecular Biophysics under the direction of Jeremy Smith at Oak Ridge National Laboratory. In 2014, he was awarded a Doctorate of Philosophy degree in Life Science with an interdisciplinary minor in Computer Science.