12-2008

# Structure- and Ligand-Based Design of Novel Antimicrobial Agents

Kirk Edward Hevener
*University of Tennessee Health Science Center*

# Structure- and Ligand-Based Design of Novel Antimicrobial Agents



**DOI**
10.21007/etd.cghs.2008.0136

# STRUCTURE- AND LIGAND-BASED DESIGN OF NOVEL ANTIMICROBIAL AGENTS

A Dissertation
Presented for
The Graduate Studies Council
The University of Tennessee
Health Science Center

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy
From The University of Tennessee

By
Kirk Edward Hevener
December 2008

## DEDICATION

This work is dedicated with love to my family: to my parents, Eugene and Ellen, my sisters Michelle and Kara, and my brother, Kevin; for their patience, support, and encouragement; to the brothers of the Omega Chapter of Phi Delta Chi, for always being there; and to Michael Cox, a truer friend one could not hope for.

## ACKNOWLEDGMENTS

I would like to express my sincerest gratitude to my mentor and research advisor, Dr. Richard E. Lee for his support, guidance, and encouragement during the course ofmy graduate studies. I also sincerely appreciate the guidance and advice of my committee members: Dr. Duane Miller, Dr. Bob Moore, Dr. Brien Neudeck, and Dr. Stephen White. I thank the members of the Lee lab for their friendship and companionship over the years and acknowledge the work of Robin Lee, Dr. Elizabeth Carson, and Dr. Julian Hurdle for performing MIC studies to measure the whole cell activity of the compounds discussed in this work. I also acknowledge the organic synthesis work of Dr. Kris Virga, Dr. Rajendra Tangallapally, Dr. Raghunandan Yendapally, and Dr. Jianjun Qi (former Lee lab members) for their efforts on design and synthesis of many of the compounds discussed in this work.

I acknowledge and thank the members of the White lab for their collaborative efforts on the DHPS research project. I would like to particularly acknowledge the work of Dr. Kerim Babaoglu, who solved the majority of the DHPS crystal structures I used in my research project. I also thank Dr. Iain Kerr and Kate Ayers for the subsequent crystallography work they performed in the DHPS project. Dr. Mi-Kyun Yun's work in implementing and running the DHPS enzyme assay was instrumental to this research and her efforts are also gratefully acknowledged.

I thank Dr. John Buolamwini in the Department of Pharmaceutical Science at the University of Tennessee for his assistance and guidance with the GOLD docking program and his valuable assistance with the 3D-QSAR studies. I would also like to express my sincere appreciation to David Ball, my friend and colleague, for his tireless assistance with this research and his help in the preparation of the manuscripts generated from the work presented herein.

Funding for this research was provided by the National Institutes of Health, ALSAC, and the University of Tennessee, College of Pharmacy. I gratefully acknowledge the American Foundation for Pharmaceutical Education for the pre-doctoral fellowship I was provided during my final two years of graduate course work as well as the UT College of Pharmacy's Feurt Scholarship and the University of Tennessee's National Alumni Association's Andy Holt Scholarship for the funding and support I received during and after my pharmacy training.

## ABSTRACT

The use of computer based techniques in the design of novel therapeutic agents is a rapidly emerging field. Although the drug-design techniques utilized by Computational Medicinal Chemists vary greatly, they can roughly be classified into structure-based and ligand-based approaches. Structure-based methods utilize a solved structure of the design target, protein or DNA, usually obtained by X-ray or NMR methods to design or improve compounds with activity against the target. Ligand-based methods use active compounds with known affinity for a target that may yet be unresolved. These methods include Pharmacophore-based searching for novel active compounds or Quantitative Structure-Activity Relationship (QSAR) studies. The research presented here utilized both structure and ligand-based methods against two bacterial targets: *Bacillus anthracis* and *Mycobacterium tuberculosis*. The first part of this thesis details our efforts to design novel inhibitors of the enzyme dihydropteroate synthase from *B. anthracis* using crystal structures with known inhibitors bound. The second part describes a QSAR study that was performed using a series of novel nitrofuranyl compounds with known, whole-cell, inhibitory activity against *M. tuberculosis*.

Dihydropteroate synthase (DHPS) catalyzes the addition of *p*-amino benzoic acid (*p*ABA) to dihydropterin pyrophosphate (DHPP) to form pteroic acid as a key step in bacterial folate biosynthesis. It is the traditional target of the sulfonamide class of antibiotics. Unfortunately, bacterial resistance and adverse effects have limited the clinical utility of the sulfonamide antibiotics. Although six bacterial crystal structures are available, the flexible loop regions that enclose *p*ABA during binding and contain key sulfonamide resistance sites have yet to be visualized in their functional conformation. To gain a new understanding of the structural basis of sulfonamide resistance, the molecular mechanism of DHPS action, and to generate a screening structure for high-throughput virtual screening, molecular dynamics simulations were applied to model the conformations of the unresolved loops in the active site. Several series of molecular dynamics simulations were designed and performed utilizing enzyme substrates and inhibitors, a transition state analog, and a pterin-sulfamethoxazole adduct. The positions of key mutation sites conserved across several bacterial species were closely monitored during these analyses. These residues were shown to interact closely with the sulfonamide binding site. The simulations helped us gain new understanding of the positions of the flexible loops during inhibitor binding that has allowed the development of a DHPS structural model that could be used for high-through put virtual screening (HTVS). Additionally, insights gained on the location and possible function of key mutation sites on the flexible loops will facilitate the design of new, potent inhibitors of DHPS that can bypass resistance mutations that render sulfonamides inactive.

Prior to performing high-throughput virtual screening, the docking and scoring functions to be used were validated using established techniques against the *B.*

*anthracis* DHPS target. In this validation study, five commonly used docking programs, FlexX, Surflex, Glide, GOLD, and DOCK, as well as nine scoring functions, were evaluated for their utility in virtual screening against the novel pterin binding site. Their performance in ligand docking and virtual screening against this target was examined by their ability to reproduce a known inhibitor conformation and to correctly detect known active compounds seeded into three separate decoy sets. Enrichment was demonstrated by calculated enrichment factors at 1% and Receiver Operating Characteristic (ROC) curves. The effectiveness of post-docking relaxation prior to rescoring and consensus scoring were also evaluated. Of the docking and scoring functions evaluated, Surflex with SurflexScore and Glide with GlideScore performed best overall for virtual screening against the DHPS target.

The next phase of the DHPS structure-based drug design project involved high-throughput virtual screening against the DHPS structural model previously developed and docking methodology validated against this target. Two general virtual screening methods were employed. First, large, virtual libraries were pre-filtered by 3D pharmacophore and modified Rule-of-Three fragment constraints. Nearly 5 million compounds from the ZINC databases were screened generating 3,104 unique, fragment-like hits that were subsequently docked and ranked by score. Second, fragment docking without pharmacophore filtering was performed on almost 285,000 fragment-like compounds obtained from databases of commercial vendors. Hits from both virtual screens with high predicted affinity for the pterin binding pocket, as determined by docking score, were selected for *in vitro* testing. Activity and structure-activity relationship of the active fragment compounds have been developed. Several compounds with micromolar activity were identified and taken to crystallographic trials.

Finally, in our ligand-based research into *M. tuberculosis* active agents, a series of nitrofuranylamide and related aromatic compounds displaying potent activity was investigated utilizing 3-Dimensional Quantitative Structure-Activity Relationship (3D-QSAR) techniques. Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA) methods were used to produce 3D-QSAR models that correlated the Minimum Inhibitory Concentration (MIC) values against *M. tuberculosis* with the molecular structures of the active compounds. A training set of 95 active compounds was used to develop the models, which were then evaluated by a series of internal and external cross-validation techniques. A test set of 15 compounds was used for the external validation. Different alignment and ionization rules were investigated as well as the effect of global molecular descriptors including lipophilicity (cLogP, LogD), Polar Surface Area (PSA), and steric bulk (CMR), on model predictivity. Models with greater than 70% predictive ability, as determined by external validation and high internal validity (cross validated $r^2 > .5$) were developed. Incorporation of lipophilicity descriptors into the models had negligible effects on model predictivity. The models developed will be used to predict the activity of proposed new structures and advance the development of next generation nitrofuranyl and related nitroaromatic anti-tuberculosis agents.

**TABLE OF CONTENTS**

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF EQUATIONS

## LIST OF ABBREVIATIONS

| | |
|---|---|
| ADME | Absorption, Distribution, Metabolism, Elimination |
| AUC | Area Under the Curve |
| AU-ROC | Area Under the Receiver-Operating Characteristic Curve |
| cLogP | Calculated Logarithm of the Partition Coefficient |
| CoMFA | Comparative Molecular Field Analysis |
| CoMSIA | Comparative Molecular Similarity Indices Analysis |
| CMR | Calculated Molar Refractivity |
| CPU | Central Processing Unit |
| DHFR | Dihydrofolate Reductase |
| DHPP | Dihydropterin Pyrophosphate |
| DHPS | Dihydropteroate Synthase |
| DNA | Deoxyribonucleic acid |
| GAFF | General Atom Force Field |
| GB/SA | Generalized Born Surface Area |
| GPCR | G Protein-Coupled Receptor |
| HIV | Human Immunodeficiency Virus |
| HTD | High-Throughput Docking |
| HTS | High-Throughput Screening |
| ITC | Isothermal Titration Calorimetry |
| Log | Logarithm |
| LOO | Leave-One-Out |
| MC | Monte Carlo |
| MD | Molecular Dynamics |
| MDR-TB | Multi-Drug Resistant Tuberculosis |
| MIC | Minimum Inhibitory Concentration |
| MM | Molecular Mechanics |
| MRSA | Methicillin-Resistant *Staphylococcus Aureus* |
| NMR | Nuclear Magnetic Resonance |
| MR | Molar Refractivity |
| MS | Mass Spectrometry |
| MW | Molecular Weight |
| NME | New Molecular Entity |
| *p*ABA | *para*-Amino Benzoic Acid |
| PB/SA | Poisson-Boltzmann Surface Area |
| PLS | Partial Least Squares |
| PME | Particle-Mesh Ewald |
| PPi | Pyrophosphate |
| PSA | Polar Surface Area |
| QSAR | Quantitative Structure-Activity Relationship |
| RHEL | Red Hat Enterprise Linux |
| RMSD | Root Mean Square Deviation |

| | |
|---|---|
| ROC | Receiver-Operating Characteristic |
| SAMPLS | Sample-Distance Partial Least Squares |
| SAR | Structure-Activity Relationship |
| SBDD | Structure-Based Drug Design |
| SEE | Standard Error of Estimate |
| SEP | Standard Error of Prediction |
| SMX | Sulfamethoxazole |
| SPR | Surface Plasmon Resonance |
| SSLR | Sum of the Sum of Log Rank |
| TB | Tuberculosis |
| VS | Virtual Screening |
| XDR-TB | Extensively Drug-Resistant Tuberculosis |
| ZINC | ZINC Is Not Commercial |
| 1D | One Dimensional |
| 2D | Two Dimensional |
| 3D | Three Dimensional |

# CHAPTER 1. INTRODUCTION

## 1.1 Computer-Aided Drug Design and Molecular Modeling

The role of computers in the design of novel therapeutic agents has a long history. As early as the 1960's computers were being used to visualize drug-target interactions. In fact, the origin of the use of computers for molecular graphics and modeling in drug discovery has been traced to Project MAC (Multiple Access Computer) at MIT in 1963.[1] By the early to mid 1970's, x-ray crystal structures of biological drug targets were being visualized and the insights gained employed in lead optimization. Over the last several decades, the exponential increase in computing power, technology, and the number of solved target structures using high-throughput X-ray, NMR, and homology modeling methods has resulted in a dramatic rise in the use of computers in many aspects of drug design. Computational techniques including quantum mechanical calculations, molecular mechanics operations, molecular simulations, graphical visualization, cheminformatics, molecular docking and quantitative structure-activity relationship studies are all being used with increasing frequency and success in the discovery and development of clinical drug candidates. These methods are being applied in almost every area of drug design, from hit identification and lead modification to metabolism, distribution and toxicology predictions. Table 1.1 lists several representative examples of clinical drugs for which computational techniques played a

**Table 1.1. Examples of Clinical Drugs Developed Using Computer-Aided Methods**

| Compound | Target | Therapeutic Use | Company | Approved |
|---|---|---|---|---|
| Captopril | Angiotensin Converting Enzyme | Hypertension | Par Pharma | 1982 |
| Saquinavir | HIV Protease[1] | HIV Infection | Roche | 1995 |
| Donepezil | Acetylcholinesterase | Alzheimer's | Eisai | 1996 |
| Nelfinavir | HIV Protease[1] | HIV Infection | Pfizer | 1997 |
| Dorzolamide | Carbonic Anydrase | Glaucoma | Merck | 1998 |
| Amprenavir | HIV Protease[1] | HIV Infection | GlaxoSmithKline | 1999 |
| Zanamivir | Neuraminidase | Influenza | GlaxoSmithKline | 1999 |
| Oseltamivir | Neuraminidase | Influenza | Roche | |
| Lopinavir | HIV Protease[1] | HIV Infection | Abbott | 2000 |
| Imatinib | bcr-abl Kinase | Leukemia | Novartis | 2003 |
| Erlotinib | EGFR Kinase | Cancers | OSI Pharma | 2004 |
| Ximelogatran | Thrombin | Anticoagulant | AstraZeneca | 2004[2] |
| Raltegravir | HIV Integrase | HIV Infection | Merck | 2007 |

1. Nearly all of the 10 marketed HIV protease inhibitors were developed using Structure-Based Computational Techniques.
2. Ximelogatran was only approved in Europe and subsequently withdrawn for high incidence of liver toxicity.

large role in their development. The compounds listed in the table were discovered and/or developed using a variety of the techniques listed above. Captopril, for example, was developed using rational drug design techniques where a homology model of the drug target, angiotensin converting enzyme (ACE), was built using the available x-ray structure of carboxypeptidase A, whose active site was hypothesized to be similar to ACE.[2,3] The HIV protease inhibitors and the neuraminidase inhibitors, used for HIV and influenza infections, respectively, were discovered (and are still being discovered) using molecular modeling and visualization techniques that utilize the solved x-ray crystal structures of their respective target enzymes.[4,5] The newly developed HIV integrase inhibitor, raltegravir, was discovered using a type of virtual screening known as pharmacophore searching, which will be discussed below.[6] Ligand-based 3D-QSAR methods were applied in the discovery efforts that led to the development of donepezil as a potent inhibitor of acetylcholinesterase for the treatment of Alzheimer's disease.[7] Finally, the kinase inhibitors imatinib and erlotinib were identified using molecular docking as the lead identification tool.[8,9] In addition to currently marketed agents, there are a large number of drug candidates in clinical trials that have been or are being investigated using computational methods; including muscarinic antagonists, somatostatin and growth hormone analogs, urotensin II antagonists, and other GPCR binding agents.[10] It is obvious that computational methods can make a large contribution to drug discovery efforts, and that as computational power continues to increase and technology advances, the role of computers in drug discovery will also continue to expand. Although there are a number of ways that computers can aid drug-design projects (as described above), the work discussed herein utilized two main computational approaches which will be described below: virtual screening and molecular simulations.

## 1.2 Virtual Screening for Lead Identification

Virtual screening (VS), as it applies to drug discovery, can be defined as the use of computational methods to discover novel compounds with activity against biological targets. It is primarily employed as a lead identification technique and has gained considerable acceptance in recent years. This compares with the traditional lead identification method of high-throughput screening (HTS), where test compounds are physically screened against the biological target at a standard concentration, usually 10 μM, using a specialized enzyme or receptor assay. Virtual screening has been shown to be a complementary tool to HTS.[11] It has several advantages over HTS as a lead identification method: First, the number of compounds that can be screened within a reasonable amount of time is much greater than HTS, on the order of $10^{15}$ versus $10^{6}$ compounds with HTS. Second, because only those compounds with predicted activity against the biological target are actually tested *in vitro,* the cost of performing a virtual screen is considerably less than that associated with HTS. Third, compounds can be built into virtual libraries for screening that have not yet been synthesized, saving the considerable time and expense of building a screening library for high-throughput screening. Finally, because VS yields a the smaller number of compounds that are

actually tested *in vitro*, the hit rates from virtual screening can be 2 to 3 orders of magnitude greater than those normally seen with HTS.[11] Virtual screening is not without its disadvantages, however.[12] It is considered an information rich method and in many cases requires structural knowledge of the target or some existing knowledge of active compounds, neither of which may be available. Although steadily improving in predictive ability, speed, and accuracy, the computational algorithms employed in virtual screening are still limited by inaccurate activity predictions. Additionally, virtual screening requires expert users to work with the programs and algorithms as well as to build and maintain the virtual compound libraries that are to be screened. Finally, in some cases the 'hit' compounds from virtual screening may not be synthetically feasible. Although HTS has advantages over VS in the areas listed above, it also is not without its disadvantages, including high cost, lower number of compounds that can be screened and lack of structural binding information for 'hit' compounds. Additionally, high-throughput screens can be troubled by frequent-hitting, false positive compounds that must be identified and eliminated.[13,14] Interestingly, studies comparing HTS and VS side by side have shown that although VS can be expected to produce higher hit rates, the hits produced are often different that those confirmed hits from the corresponding high-throughput screens.[11,15] This may imply that rather than acting as competing methods of lead identification, high-throughput screening and virtual screening should be considered complementary methods and used together to identify and test promising leads.

Although there are numerous methods for performing a virtual screen, they can be roughly classified into two main types: ligand-based approaches which do not utilize the structure of the biological target in screening, and docking-based approaches, which utilize the structure of the biological target, usually obtained by NMR or x-ray methods, and a variety of molecular docking algorithms and scoring functions. Hybrid approaches which combine aspects from ligand-based and docking-based methods are also frequently employed in virtual screening studies.

### 1.2.1 Ligand-Based Approaches

Ligand-based approaches typically utilize knowledge of a set of compounds with known activity against the biological target. These approaches are frequently employed in the absence of structural information on the target in question and in addition to lead identification, can also be used as a lead modification strategy. The key concept in ligand based approaches is that compounds that are structurally similar or have similar structural components to the known active compounds are more likely to have activity themselves. A variety of ligand-based screening methods have been developed that are being used with increasing frequency, such as substructure searching, similarity searching, pharmacophore searching, clustering methods, and finally QSAR and 3D-QSAR methods. Each of these methods is similar in that they utilize the structural features of known active compounds, but they differ in their computational requirements, search algorithm, and the features of the hits compounds they return. Interestingly, several of the methods discussed below have been successfully used in 'lead-hopping'

or ‘scaffold-hopping’ studies, which reflect the ability of ligand-based approaches to identify lead compounds outside the structural class of the known active compounds upon which the screen was based.[16-22] The methods listed above can also be classified by whether they employ 2-dimensional (2D) or 3-dimensional (3D) methods for searching the virtual compound databases. 2D methods utilize the chemical structure of the active compounds, whereas 3D methods incorporate the 3-dimensional shape of the active compounds in addition to the chemical structure and/or structural features. 3D methods make assumptions as to the binding conformation of the known active compounds.

Substructure searching is a relatively simple screening method that performs a search of a compound database to match a specified structural feature, i.e. functional group, ring system, etc. Typically, in compound databases, the structural features are represented by searchable bitstrings, or binary representations. Two types of bitstrings are commonly used, structural keys and hashed fingerprints. In a structural key, every position in the bitstring represents a particular structure. The structural key utilizes a fragment dictionary and assigns a 0 if the structure is absent from the compound and a 1 if the structure is present. Structural keys are easily and quickly searched but require the added structure library and recalculation of bitstrings when compounds are added to the compound database with new structures. A hashed fingerprint bitstring does not use a fragment dictionary; instead an algorithm is employed to assign bits to specify given structural patterns in a compound. All possible linear paths of atom connectivity are calculated up to a predefined number of atoms (typically 8) and bits are assigned to represent each path. Each pattern may require several bits to be represented. For example glyceraldehyde, shown in Figure 1.1, contains the following paths of length 4: O-C-C-O, O-C-C-C, O-C-C=O, and C-C-C=O. Each pattern (or path) would be assigned a unique, searchable set of bits which are each set to a value of 1. A given compound can be represented by bitstrings of up to several thousand bits after all atom paths have been calculated.

While substructure searching is a useful and quick method of searching a compound library for 2D features, it does not take into account 3-dimensional conformations of the compounds or physicochemical properties of the atoms or functional groups being searched. Pharmacophore searching can be considered a special type of 3D substructure searching that in addition to the 3D conformations, can also take into account functional group features such as polarity, hydrogen bond potential, aromaticity, and hydrophobicity. An added advantage to pharmacophore

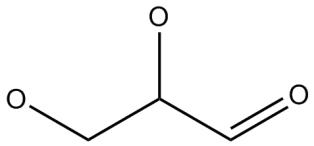



**Figure 1.1. Glyceraldehyde with Hydrogens Suppressed**

searching is that, unlike 2D substructure searching, this method can identify lead compounds that are structurally dissimilar to those already known, a process known as 'scaffold-hopping' or 'lead-hopping'. In pharmacophore searching, a set of features common to the known active compounds is identified and used for the search criteria. A 3D pharmacophore search can include structural fingerprints, 3D spatial constraints and 'macros' which define the physicochemical properties for substructures (H-bond donor or acceptor, hydrophobic, etc.). Figure 1.2 shows an example of a defined 3D pharmacophore search using a DHPS product analog. Three key features have been defined: an aromatic center, a hydrogen bond acceptor, and a hydrogen bond donor atom. In order to match these criteria, a searched compound must not only contain the three features specified, but also in the correct 3D alignment. In addition to searching 3D space, some advanced search algorithms are even able to modify torsional angles of compounds being searched to test whether the compound can adopt the specified pharmacophore alignment, a process commonly called flexible searching. It is also common to use constraints to limit the number of compounds being searched and reduce computational expense. Constraints can include simple drug-like criteria such as Lipinski or Veber rules for molecular weight, numbers of rotatable bonds, and other key features.[23,24] They can also incorporate known structural information of the biological target's active site, if any is known, in the form of exclusion spheres or a molecular surface, both of which create barriers the compounds being searched are not permitted to encroach.

Similarity searching is another popular method of identifying compounds with similar structural features to the known active compounds. It differs from substructure searching and pharmacophore searching in that there is no precise query that the molecule being searched can match. This search technique involves calculating and comparing similarity coefficients between the known active compound and the compounds being screened. The similarity coefficients can be based upon any number of molecular descriptors. The compounds which score the highest in the similarity search are considered the hit compounds and theoretically would be tested for biological activity. Some common molecular descriptors that have been used in similarity searches include molecular weight; hashed fingerprints and structural keys; counts of atoms, rings, or other features; octanol/water partition coefficient; molar refractivity; molecular connectivity ($\chi$) indices; shape ($\kappa$) indices; electrotopological indices; atom pairs and topological torsions; dipole moment; molecular volume, surface area, or polar surface area; quantum chemical descriptors (HOMO, LUMO, energies, etc.); partial atomic charge and polarizability; pharmacophore keys; and geometric atom pairs, torsions, and angles.[25] After the appropriate molecular descriptors have been calculated for the compounds of interest, similarity coefficients are calculated to make the comparison. These coefficients can be calculated using one of several different methods, with probably the most common being the Dice coefficient, the Cosine coefficient, and the Tanimoto coeffiecient.[25] The Tanimoto coefficient is commonly used for binary data (structural keys, fingerprints, etc.) and the formula is given below in

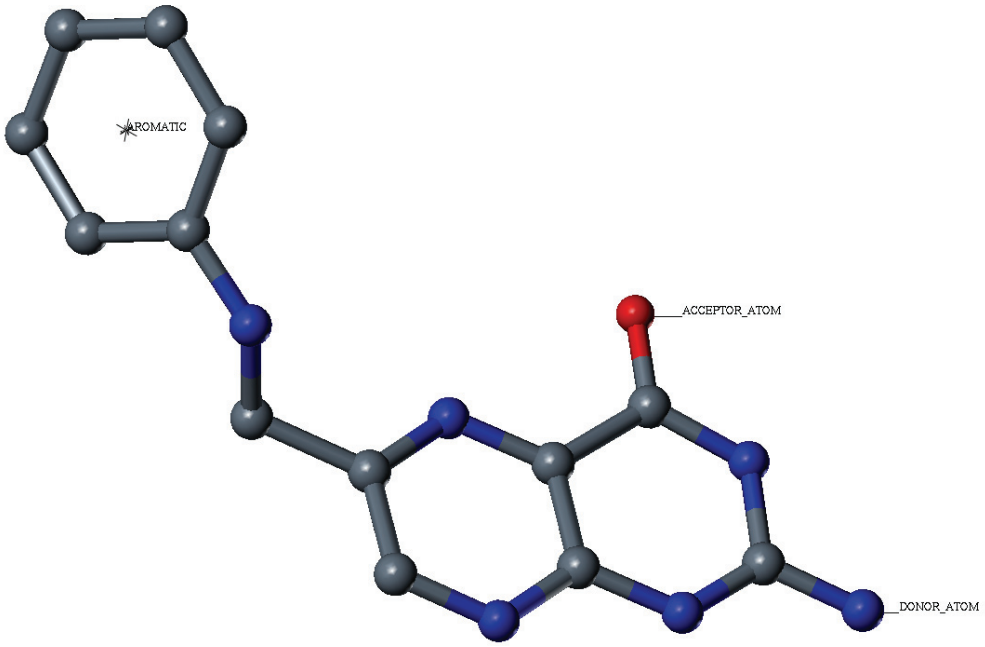



**Figure 1.2. A Ligand-Based 3D Pharmacophore Search**

Equations 1.1 and 1.2 for continuous variables and binary variables, respectively. Similarity can be calculated based on 2D as well as 3D descriptors. Similarity measurements using 2D descriptors will generally associate molecules with similar substructures, while those using 3D descriptors are able to account for 3D pharmacophore and molecular recognition and binding. The advantage of using 3D descriptors, therefore, is the increased potential for locating active compounds with unique scaffolds.

$$S_{AB} = \frac{\sum_{i=1}^{N} X_{iA} X_{iB}}{\sum_{i=1}^{N} (X_{iA})^2 + \sum_{i=1}^{N} (X_{iB})^2 - \sum_{i=1}^{N} X_{iA} X_{iB}} \qquad \textbf{Equation 1.1}$$

$$S_{AB} = \frac{c}{a + b - c} \qquad \textbf{Equation 1.2}$$

Occasionally it may be desirable to select a set of diverse compounds from a library for screening. This is often done in order to decrease the number of compounds being tested while still sampling the maximum diversity of the screening library. In this case, a method known as cluster analysis or clustering can be very helpful. Clustering utilizes essentially the same methods as similarity searching, with the exception that compounds are selected based upon dissimilarity. In a cluster analysis, groups of similar compounds (clusters or bins) are created from which representative compounds can be selected. There are a number of clustering methods in popular use today, the most common being Jarvis-Patrick clustering and Hierarchical clustering.[26,27] Hierarchical clustering seems to have outperformed Jarvis-Patrick clustering in terms of predicting property values and activity by cluster placement in two recent studies and of the two, is the more popular clustering method.[28,29] Other methods for selecting dissimilar compounds include dissimilarity-based methods and partition-based methods. Because neither of these methods were utilized in the work described here, we will not expand on them further.

The last and probably most frequently utilized ligand-based method that will be discussed is the quantitative structure-activity relationship, QSAR. QSAR techniques are methods used to correlate physicochemical descriptors from a set of related compounds to their known molecular activity or molecular property values. QSAR models can be very useful in predicting the activity of compounds which have not been tested *in vitro* and are frequently used in virtual screening to identify lead compounds as well as to prioritize synthetic efforts. The first QSAR studies are usually attributed to Hansch and coworkers, who correlated biological activity of a series of compounds with their hydrophobic and electrostatic properties.[30] There are a variety of descriptors that have been used to develop QSAR models, many of which were listed above in the similarity searching discussion. They can range from connectivity and shape descriptors to molecular descriptors for lipophilicity (cLogP and LogD)[31,32], steric bulk (Molar Refractivity, volume)[33], and electrostatics (polar surface area, Coulombic charges, dipole moments).[34]

In addition to 2D and global molecular descriptors, QSAR models can be built using 3-dimensional molecular descriptors. Most 3D-QSAR models require the alignment of the active compounds into a known or theoretical binding conformation. There are several different methods that are used for the calculation of 3D descriptors, the most common being comparative molecular field analysis (CoMFA)[35] and comparative molecular similarity indices analysis (CoMSIA).[36] CoMFA involves the calculation of steric and electrostatic values using charged probe atoms at grid lattice points while CoMSIA utilizes 3-D similarity indices. Other methods used to calculate 3D descriptors include comparative molecular moments analysis (CoMMA),[37] a molecular vibration-based method (EVA),[38] weighted holistic invariant molecular indices (WHIM),[39] and hypothetical active site lattice (HASL).[40] 3D-QSAR methods have an advantage over traditional QSAR in that they can provide information regarding the nature of the biological target's active site, in terms of favorable binding regions and characteristics, which can be very useful to the drug design efforts.

Once the QSAR descriptors have been calculated, the QSAR equation is derived using a regression tool that is applicable to the data being utilized. The independent variables (descriptors) are used to derive the equation that predicts the dependent variable (activity or property). For 2D QSAR models a simple linear regression or multiple linear regression (MLR), if there are several independent variables, is usually sufficient. MLR cannot be used for 3D-QSAR models where there the number of independent variables greatly exceeds the number of dependent variables (i.e. the number of compounds in the training set). In these cases, one of two methods are commonly used, principal components regression (PCR) or partial least squares (PLS).[41] In PCR the independent variables are subjected to a principal components analysis, after which a regression is performed using the first (usually 2 or 3) principal components. Validation methods (described below) can help the model developer determine the appropriate number of components to use in the final model. PLS uses linear combinations of the independent variables to describe the dependent variable. A sample PLS equation follows:

$$y = b_1 t_1 + b_2 t_2 + b_3 t_3 + \ldots b_m t_m \qquad \textbf{Equation 1.3}$$

where *y* is the dependent variable, $b_m$ is a calculated coefficient, and

$$t_1 = c_{11} x_1 + c_{12} x_2 + \cdots + c_{1p} x_p \qquad \textbf{Equation 1.4}$$

The latent variables (or components) in a PLS analysis are the *t* values, calculated by linear combination of the independent variables (x). PLS is different from PCR in that it can explain variations not only in the dependent variables, *y*, but also variation in the independent variables. The number of latent variables used in the final QSAR model is again determined by a variety of validation methods. Other methods for deriving QSAR equations include discriminant analysis, neural networks, and inductive logic programming.[42,43]

Once a QSAR model has been built, it must be validated prior to use by a variety of internal and external validation methods. The goodness of fit of the QSAR equation is usually given by the multiple correlation coefficient, $r^2$, when deriving the equations. Values close to unity are desirable and indicate a high degree of internal validity. The F statistic and Standard Errors of Estimate (SEE) are also commonly used to validate the goodness of fit. Cross-validation methods include the commonly used Leave-One-Out (LOO) and Leave-Group-Out (LGO) methods. In cross-validation, one or more of the training compounds (known actives used to derive the QSAR model) are left out during model derivation and then the model derived is used to predict the activity of the compound or compounds left out. This process is normally repeated many times and a mean $q^2$ (cross-validated $r^2$) value is determined. The $q^2$ value is a measure of goodness of prediction of the QSAR model derived. Cross-validation is considered in internal validation because it used training set compounds to generate the $q^2$ value. One of the most rigorous methods for validating QSAR models is known as external validation, or test set validation, where compounds with known activity that were not used in creation of the QSAR model are used for activity predictions and $r^2$ values are derived from these predictions. Finally, bootstrapping is a method that can be used to obtain confidence intervals for the $r^2$ and SEE values.

### 1.2.2 Docking-Based Approaches

Knowledge of the biological target's structure, in particular the targeted binding site, is most desirable from a drug discovery perspective because direct knowledge of ligand binding interactions can be utilized in drug design efforts; a procedure that has come to be called 'Rational Drug Design' or 'Structure-Based Drug Design' (SBDD). Target structures are usually obtained by solving an x-ray crystal or NMR structure, although homology modeling methods are also sometimes employed. There are a variety of SBDD methods that can be used, the most common being de novo design and molecular docking.

De novo design uses the 3-dimensional structure of the target's active site to guide the placement and linking of molecular fragments obtained from fragment databases. There are two general methods of de novo design. In the first compounds are selected based on observed or theoretically favored binding groups determined by active site analysis, built and then placed into the active site for binding energy calculation. In the second, the fragments are placed and linked directly in the active site using build and grow strategies. Theoretically, de novo design will yield novel, active compounds which are not already present in corporate or commercial databases.

Molecular docking is a specialized form of virtual screening in which compounds are placed in the active site using a variety of search algorithms and then binding affinity is estimated using one of a number of different types of scoring functions. Requirements for molecular docking include a 3D representation of the active site, a library of compounds in a recognized electronic format, and a validated docking algorithm and

scoring function. Following docking and scoring, high scoring compounds are usually selected for testing in an *in vitro* binding assay. In recent years, molecular docking has become a very common method of virtual screening due to the exponential increase in searchable 3D structures of biological targets and improvements in computational power and technology. Figure 1.3 shows the dramatic rise in protein structures that are available at the Protein Data Bank over the last 20 years.[44] In 2007, there were over 800 articles published relating to molecular docking studies.[45] The most common docking programs in use today are shown in Figure 1.4.[46-65] The percentage of citations is shown for the most common programs (over 5 citations), determined by SCOPUS[66] search considering the original references and limited to articles in 2007.

Molecular docking involves two main processes: pose prediction and scoring. In pose prediction a search algorithm determines an optimal conformation and orientation for a given compound in the receptor, or active site. This is followed by scoring to determine whether the pose will be accepted or rejected. Generally, docking algorithms use scoring in two ways, the first is for pose selection and often uses a more simplified and rapid scoring method. The second use of scoring is when the final selected poses for all the compounds tested are scored for ranking purposes. This is often performed by a more advanced scoring function and may be computationally more intense than the pose selection scoring. Historically, there are two general types of molecular docking: rigid body docking, where the compounds are placed into the active site "as is" so to speak, normally in a minimum energy conformation; and flexible docking, which test multiple conformations of the compounds being docked. Although flexible docking is computationally more expensive, the results generated are much more accurate and this method has become the preferred method of performing molecular docking.

There are three main types of flexible ligand docking algorithms that are currently is use: systematic docking algorithms, random or stochastic algorithms, and simulation methods.[67] Table 1.2 gives a breakdown of these search methods and some representative examples of programs employing these methods. Systematic search algorithms attempt to explore all the degrees of freedom of the compounds being docked, and normally utilize one of three methods: conformational searching, fragmentation, or database methods. In conformational searching, all degrees of freedom of the compound being analyzed are explored by systematically modifying torsion angles of rotatable bonds in the compound by predefined increments. This method is very computationally expensive and is therefore rarely used. Fragmentation is probably the most popular form of flexible docking. This method breaks the compound being docked into fragments and then joins them in the active site, recreating the ligand in an energetically preferred conformation. This procedure has been called the "place-and-join" method. Alternatively, a 'core' fragment can be initially docked and then flexible sections added incrementally. This is called an "incremental construction" method. Database methods are the third type of systematic docking algorithm. In this approach, conformation libraries (or ensembles) are generated for each compound being docked and then rigidly docked.

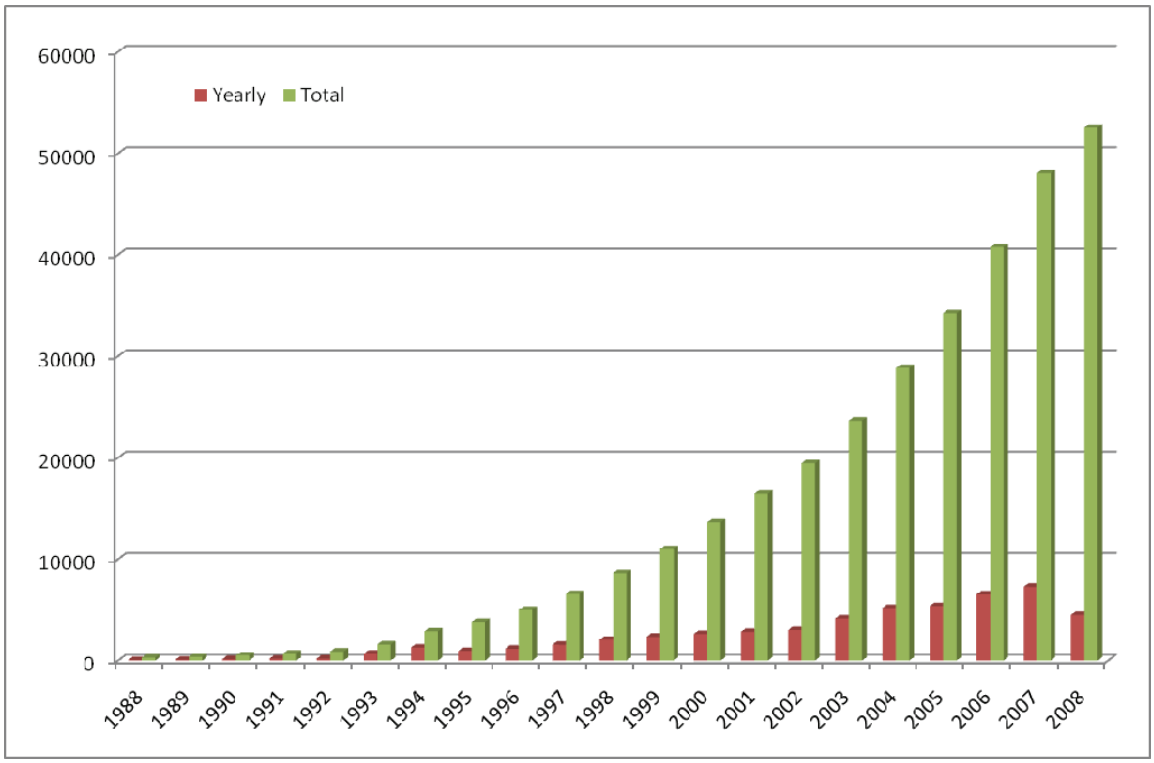



**Figure 1.3. Yearly Protein Data Bank Content Growth[44]**

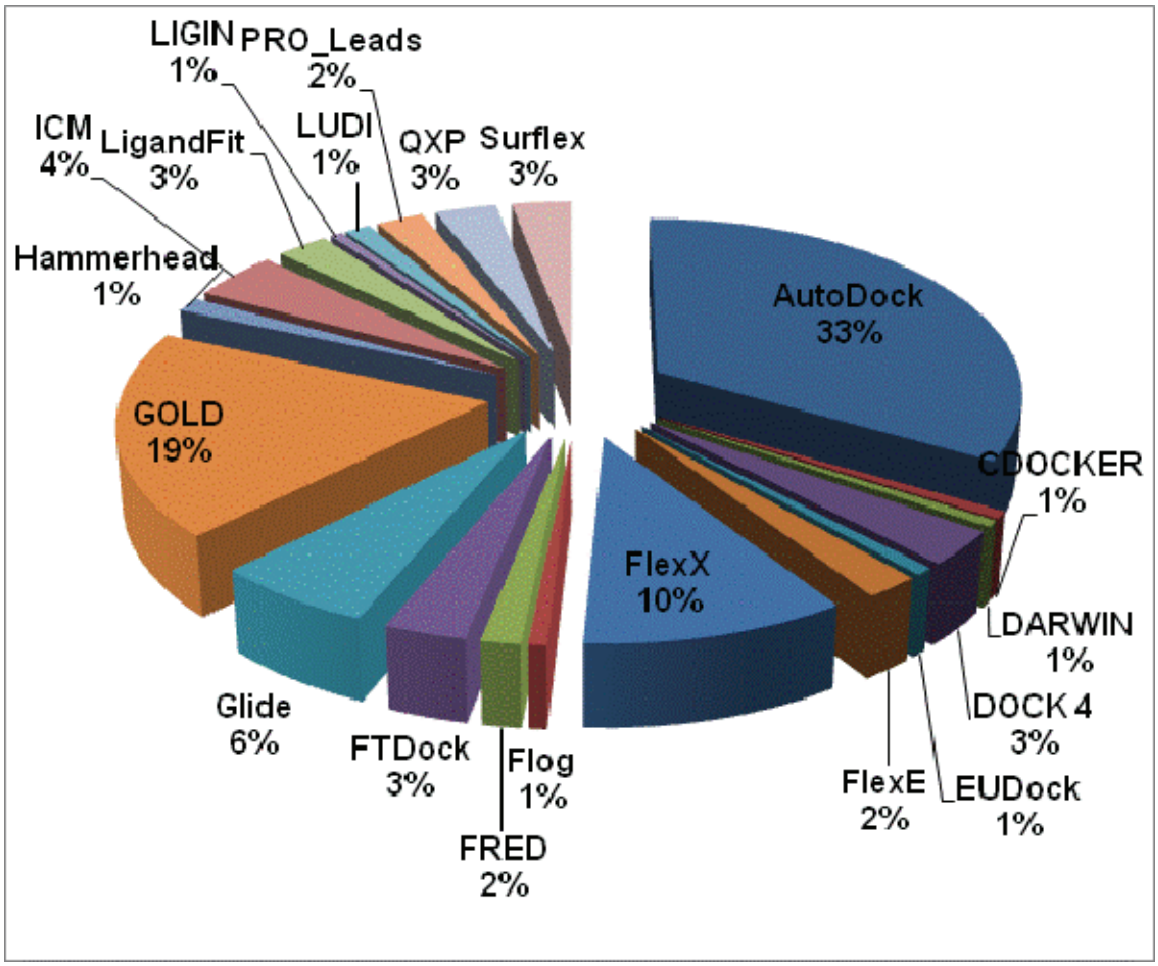



**Figure 1.4. Docking Programs by 2007 Citation**

**Table 1.2. Flexible Docking Methods and Examples**

| Methods | Representative Examples |
|---|---|
| *Systematic Docking Algorithms* | |
| Conformational Search | |
| Fragmentation | LUDI,[62] FlexX,[52] DOCK,[49] ADAM,[68] Surflex[65] |
| Database Methods | FLOG[53] |
| *Random/Stochastic Algorithms* | |
| Monte Carlo Methods | ProDock,[69] ICM,[59] MCDOCK,[70] QXP[64] |
| Genetic (Evolutionary) Algorithms | GOLD,[57] AutoDock,[46] DIVALI,[71] DARWIN[48] |
| Tabu Searching | PRO_LEADS[63] |
| *Simulation (Deterministic) Algorithms* | |
| Molecular Dynamics | Amber,[72] CHARMM,[73] NAMD,[74] GROMACS[75] |
| Minimization Techniques | Fletcher-Reeves, Newton-Raphson, Marquardt |

Random or stochastic search algorithms generate random ligand conformations (or random conformational changes), which are then docked and scored and accepted or rejected based upon predefined criteria. Random methods have an advantage over deterministic methods in that energy barriers can be bypassed, which can allow for a more complete search of conformational space. The three most popular random methods are Monte Carlo methods, Genetic Algorithms, and Tabu Searching. The criterion for accepting or rejecting poses generated in the Monte Carlo method is based on a Boltzmann probability function. Genetic algorithms utilize evolutionary techniques to generate successive "generations" of compound poses. Ligand conformations and orientations are defined by a set of variables (genes) and genetic operations perform a series of mutations, crossovers, and migrations to generate new generations which are accepted or rejected based upon a predefined fitness function. Successive generations are optimized until a final generation is determined. Tabu search methods force the search algorithm in new directions by imposing restrictions that prevent searching areas of conformational space that have already been explored. The acceptance or rejection of new poses generated is determined by RMSD calculations with a library of poses already generated.

The last main type of flexible docking is the simulation methods, which include molecular dynamics and minimization techniques. Molecular Dynamics (MD) methods work by integrating Newton's laws of motion to produce a trajectory that simulates how the system in question, in this case a ligand bound into an active site of a target biomolecule, behave over time. Dynamics methods can be very computationally expensive, but they have an advantage over many of the methods discussed above in that protein flexibility and induced fit can be taken into account. Molecular Dynamics will be discussed in detail below as it applies to simulating protein movements and dynamic structure. When MD is applied to molecular docking, typically the target macromolecule, with the possible exception of active site residues and flexible loops near the active site, is held rigid to minimize the computational expense.

The last type of search algorithm that will be discussed is energy minimization. These technique involve modifying the structure of the ligand bound in the active site to minimize the binding energy, as calculated by a variety of methods including direct searches (simplex), gradient (steepest descent), conjugate-gradient, second-derivative, and least squares methods.[76] Minimization can typically find local energy minima very well, but are not able to overcome barriers to locate global minima. These techniques also have difficulty in cases, not so uncommon, that the ligand binds to the active site in a high energy state. Minimization techniques are rarely uses as stand-alone docking methods, but they are often incorporated with other search methods in multi-step docking algorithms, for example the program Glide, which utilizes Monte Carlo and Minimization methods.[56]

Once poses are selected, scoring functions are utilized to rank the compounds by their predicted affinity for the target site. Table 1.3 lists several of the common types

**Table 1.3. Scoring Methods and Examples**

| Methods | Representative Examples |
|---|---|
| Force-Field Based Functions | D-Score,[77] G-Score,[77] GoldScore,[78] AutoDock 3.0[46] |
| Empirical Scoring Functions | LUDI,[79] F-Score,[52] ChemScore,[80] Fresno[81] |
| Knowledge-Based Functions | PMF-Score,[82] DrugScore,[83] SMoG-Score[84] |
| Consensus Functions | Cscore,[85] X-Cscore[86] |
| Solvation-Based Functions | HYDREN,[87] GB/SA,[88] SEED,[89] ZAP,[90] PB/SA[91] |

of scoring functions utilized in molecular docking, along with some representative examples of each type. Scoring functions can be generally classified into three main types: force-field based, empirical, and knowledge-based. Additionally, consensus scoring which involves combinations of scores from different functions and solvation scoring which, as the name implies, takes solvation/desolvation into account when generating a score.

Force fields (which are also called molecular mechanics) can be considered functions which calculate the energy of a system as a function of atomic positions. Force fields ignore electronic effects and typically contain bonded terms for bond stretching, angle bending, and torsion rotation and non-bonded terms for van der Waals and electrostatic interactions. Force fields and their energy terms will be described in more detail in the molecular simulations section below. Force field based scoring functions typically generate a score based upon two calculated energy values: the internal energy of the ligand and the interaction energy between the ligand and the receptor. Traditionally, force fields ignore entropic and solvation effects, which can be considered a limitation.

Empirical scoring functions are designed and trained using experimental binding energies that have been calculated from known ligand-receptor complexes. They can consist of a variety of energy terms which use coefficients determined by regression analysis of the training set binding energies. An advantage of empirical scoring functions is their ease of low computational requirement. Disadvantages include the requirement on a experimental training set and the non-transferability of the energy terms due to the parameterization process. Knowledge-based scoring functions are designed to reproduce experimental binding conformations, in contrast to empirical functions which are trained to reproduce binding energies. They generally use simple atomic interaction-pair potentials which are based on their frequency of occurrence in the training set ligand-receptor complexes used.

As mentioned above, consensus scoring combines information from several scoring functions to generate a consensus score.[85,92] The use of consensus scoring can theoretically compensate for errors in a single scoring function and improve the likelihood of identifying a correct pose. The disadvantage to consensus scoring is that systematic scoring errors can be compounded when scoring functions are correlated and this can lead to amplification of error rather than error compensation.

Many advanced scoring functions include terms that can take into account entropic and solvation/desolvation effects, both of which have been traditionally ignored in early generation scoring functions. Penalty functions that take into account the number of rotatable bonds in a compound being docked are a simple means of taking into account entropic effects. For example, the ChemScore function contains an explicit energy term for rotational energy that is intended to, in part, account for entropic effects. The effects of solvation can be accounted for in several different ways, with varying degrees of computational intensity.[93] One method that is used in force field scoring functions is the modification of the dielectric constant in the electrostatic energy term to account for the effects of solvation. Additional H-bonding terms can also be used to account for donor-water and acceptor-water effects. Buried polar and ligand desolvation energy terms have been used with success in some in empirical scoring functions, for example Fresno.[81] A more computationally expensive method is the use of a generalized-Born/surface area (GB/SA) approach, which has been successfully employed with the DOCK program.[88] Finally, a very rigorous and very computationally expensive method, Poisson-Boltzmann surface area (PB/SA) solvation scoring, has been reported.[93]

As previously mentioned, most of the search algorithms (docking programs) discussed above treat the receptor into which the compounds are being docked as a rigid body. This is one of the caveats of molecular docking studies; they do not take induced fit of the macromolecule into account. Although they are usually more computationally intense, there are a number of approaches that have been utilized in recent year to account, in some manner, for protein flexibility in docking including: molecular dynamics, Monte Carlo methods, rotamer libraries, protein ensembles, and soft receptor modeling. With the molecular dynamics and Monte Carlo methods, flexible regions of the protein can be defined for the docking runs. Unfortunately the time needed to dock a ligand will increase exponentially as the amount of flexible region is expanded, up to several days in some cases of fully flexible targets.[94] Alternatively, using rotamer libraries for side chains can represent some protein flexibility and induced fit, usually in the active site, and is less time-consuming.[95] This method, however, does not account for large protein movements. Another method is the use of an ensemble of protein conformations, calculated by a variety of methods, into each of which is docked the compounds in the screening library. This multiplies the docking time required by the number of protein conformations in the ensemble but is significantly less time consuming than the more rigorous dynamics and Monte Carlo methods. FlexE, derived from the FlexX docking program, is a popular ensemble method.[51] Finally, the soft receptor

modeling technique combines different protein conformations into one "weighted average" that is used for ligand docking.[96,97] This typically leads to enlargement of the active site as mutually exclusive binding areas are simultaneously considered, which can be considered a disadvantage. Additionally, soft receptor modeling, like rotamer libraries, cannot account for large scale protein movements.

Although we have separated virtual screening into ligand-based and docking-based methods, it should be noted that large scale virtual screening projects can incorporate aspects of both. For example, in order to decrease the time required to screen a large library of compounds, ligand-based methods such as pharmacophore searching can be utilized as filters prior to high-throughput docking. Alternatively, ligand-based methods can be used to post-process docking output in order to decrease the number of compounds requiring *in vitro* testing. An example would be applying clustering methods to docking output to select a diverse set of compounds for testing. An example of such a hybrid virtual screening workflow is shown in Figure 1.5. Virtual screening steps are listed with descriptions that include the computational intensity at each step as well as the number of screening compounds that can be handled at each stage in a reasonable amount of time; in this case the entire procedure can be completed in approximately one to two weeks for a screening library that initially numbers in the millions.

### 1.2.3 Compound Selection for Virtual Screening

The discussion on virtual screening is not complete without mention of selection processing of compounds prior to screening. The creation of a virtual screening library is a multistep process that must take into account many factors including the nature of the screening target, the desired physicochemical properties of the screening library, the time available for virtual screening, and even the type of experimental assay that will be used to test hit compounds.[98] The steps involved and considerations necessary at each step are described below.

The first step is identifying the compounds to be placed in the virtual screening library. Typically corporate or commercial libraries are screened, but the compounds can also be created *in silico* using a variety of virtual library enumeration protocols. Usually, one starts with 2D structural files, sdf or SMILES formats are most common. Once the compound files are obtained they must be analyzed and cleaned. Compounds represented in salt forms must be corrected and the salts removed, this is commonly known as desalting and a variety of algorithms are available to accomplish this. Filters can be applied at this step as well to remove unwanted compounds containing such features as reactive functional groups, unstable or hydrolizable groups, and cytotoxic groups. Figure 1.6 shows several examples of such undesirable compounds. Additionally, broken or incomplete structures as well as structures containing metals are often removed at this step. Finally, depending on the nature of the desired screening library, methods can be applied to filter the screening library for diversity or to build a

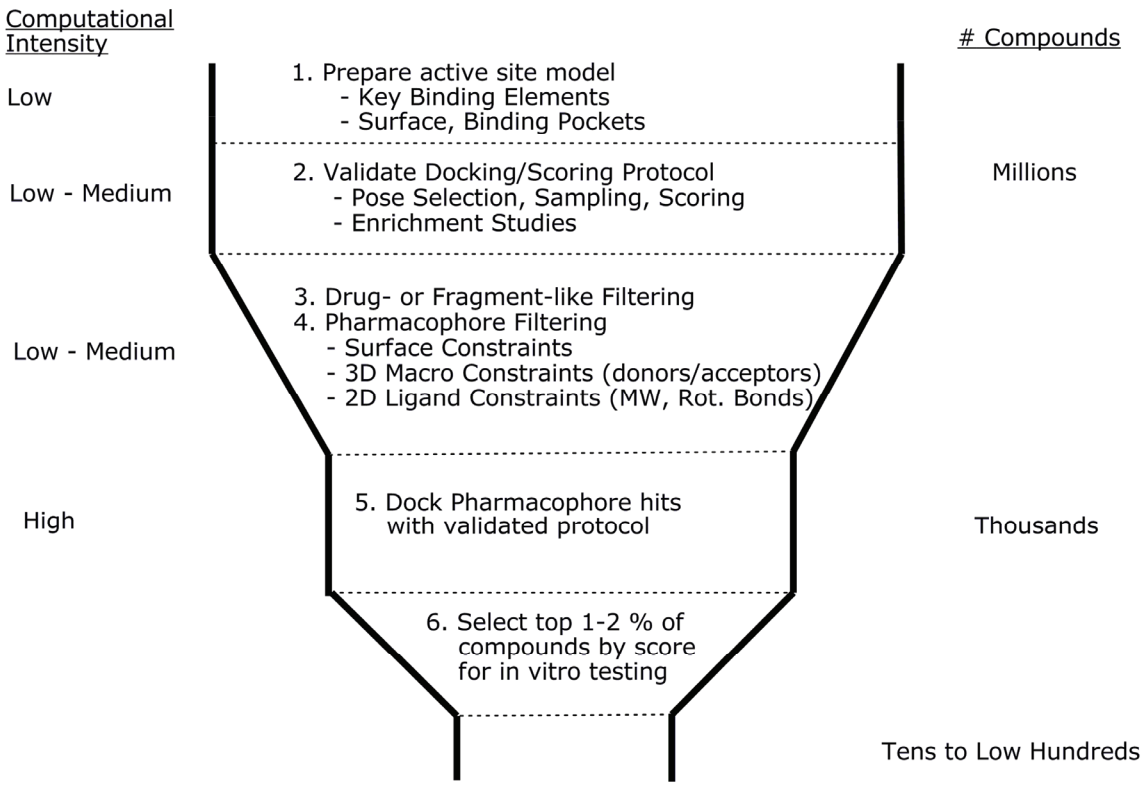
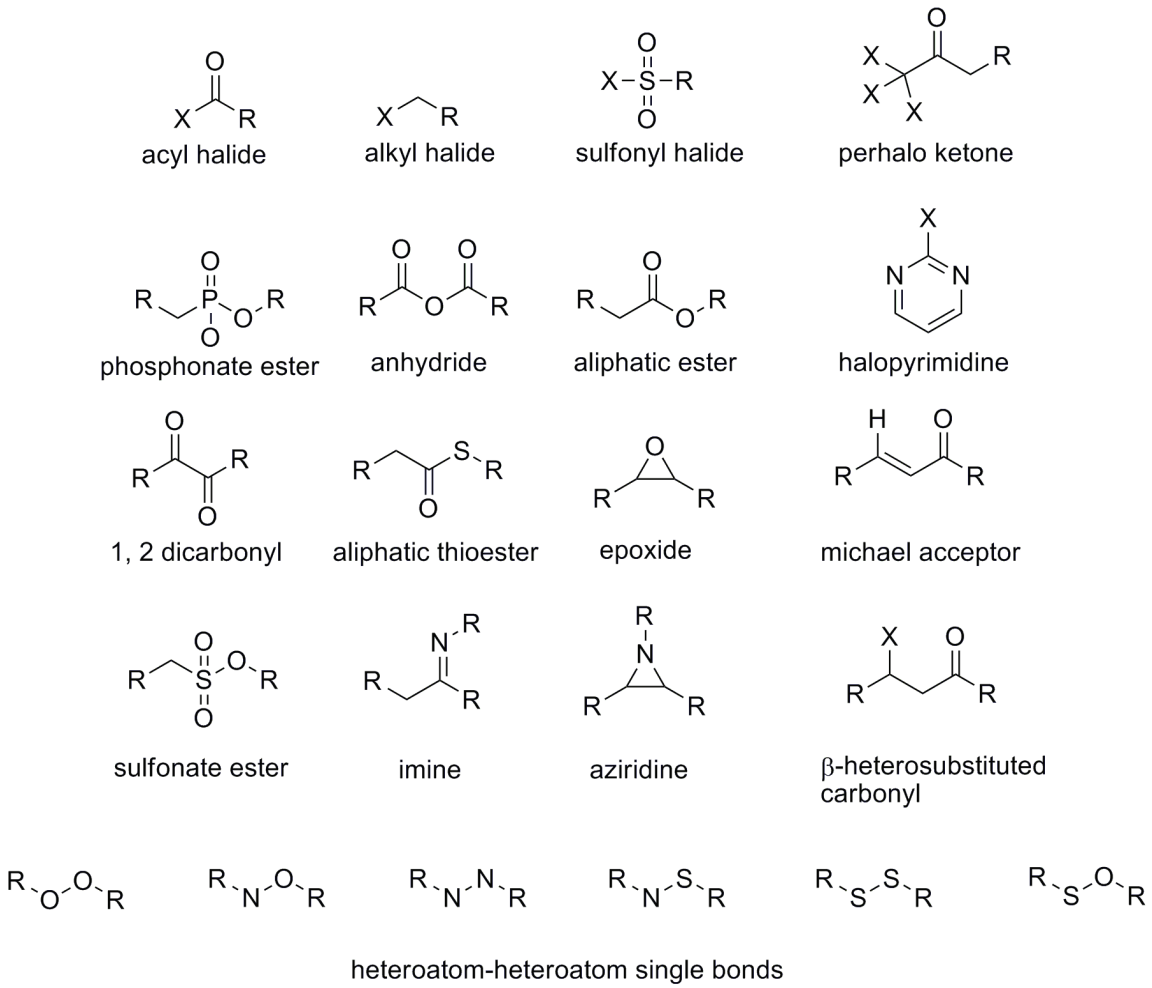



**Figure 1.5. A Docking-Based Virtual Screening Workflow**

## Reactive or Unstable Functional Groups




## Cytotoxic Functional Groups

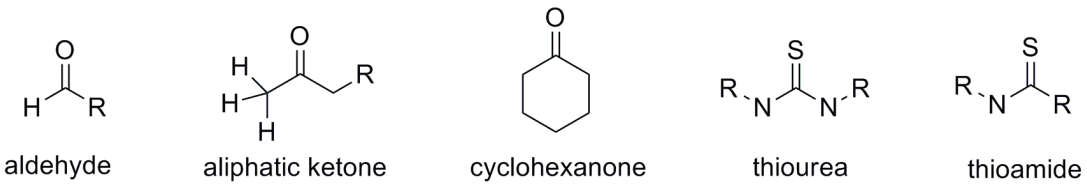



**Figure 1.6. Characteristic Undesirable Functional Groups in Virtual Screening Compounds**

Modified with permission from Rishton, G. M. Reactive compounds and in vitro false positives in HTS. *Drug Discov Today* **1997,** 2, 382-384.[13]

focused library retaining only compounds similar to known actives or containing key pharmacophoric features. Screening libraries can also be filtered using lead-like or drug-like filters including molecular weight, rotatable bond counts, ring counts, halogen counts, and h-bond donor and acceptor counts.

The next step is the generation of 3-dimensional structures from the 2D structural files. There are a variety of tools available to the molecular modeler for the generation of 3D structures from 2D structure files.[54,99,100] The two most common programs are Concord and Corina.[101,102] At this step the modeler must make choices regarding generating multiple conformations per compound, generating conformations for multiple tautomeric states, expanding compounds to account for chiral centers and taking into account protonation/deprotonation at specific pH ranges. These decisions will be based on the type of virtual screen that is to be performed as well as the computational resources available to the modeler. Depending on the detail of the 3D library that is desired, this step can result in an exponential explosion in the number of screening compounds in the virtual screening library.

Following 3D structure generation, the last step is normally loading partial atomic charges on the compounds. The size of the library will normally determine the type of charge calculation method that the modeler will choose to accomplish the procedure. The most accurate method would be to use quantum mechanical methods for charge calculation, unfortunately this very computationally expensive and is usually too time-consuming, even for small libraries. For smaller screening libraries, semi-empirical methods such as the PM3 method available in the MOPAC suite would be the most accurate; however this method can also be time consuming taking several seconds to minutes for a compound, depending on the size and complexity of the compound.[103] A variety of rules-based methods calculate charges based upon atom types, bonding and free valences and are very quick and easily implemented for very large virtual screening libraries. The disadvantage being that they do not utilize electronic calculations for calculating the atomic charges and for compounds with complex electronic systems (pi delocalization, internal h-bonding, strong electron donating or withdrawing functional groups), the atomic charges generated can be less than reliable. Some commonly used rules-based charge calculation methods are Del Re charges,[104] Gasteiger/Marsili charges,[105] Hückel charges,[106] Pullman charges,[107] and MMFF charges.[108] Once the libraries have been generated and the charges loaded, they are typically saved in a commonly used molecular file format for virtual screening, usually 3D sdf or multi-mol2 files.

## 1.3 Molecular Simulation Methods

### 1.3.1 Introduction to Simulation Methods

Molecular simulations are a way to visualize a system by generating successive configurations of the system. There are several advantages of using simulation methods in drug discovery and design. Biological systems, for example proteins, can be simulated under special conditions such as solvated and at different temperatures and pressures and with different substrates bound into the active site. This is beneficial in that, while techniques such as x-ray crystallography can generate a snapshot of a protein (or other macromolecule), the positions of mobile elements, such as flexible loops, may remain unclear. It is possible to visualize these mobile elements with simulations. Additionally, x-ray crystallography and NMR methods are often employed under non-physiological conditions (temperature, pressure, pH, solvent, etc.), which can affect their results in unpredictable ways. Simulations methods, although time consuming and computationally expensive, can provide information to the molecular modeler about how a biological system behaves over a certain time period, under physiological conditions.

There are two main types of simulation methods that will be discussed here: Molecular Dynamics (MD) and Monte Carlo (MC) simulations. With MD, the configurations are produced by integrating Newton's laws of motion, resulting in a trajectory that specifies how the system behaves with time. In this "deterministic" method, any future configuration of the system can be predicted from its current state by calculating energy and velocity for the atoms in a system using very small time steps, usually on the order of femtoseconds. The forces on the atoms are used with their current positions and velocities to predict new positions and velocities for the next time step. Over a given time period, a "trajectory" is generated that describes how the system being studied changes over time. Time averages for thermodynamic properties such as internal energy, heat capacity, pressure, and temperature can be calculated.

Monte Carlo simulations differ from dynamics in that the configurations are generated using a random approach where each configuration depends only upon the previous one. Special algorithms based upon Boltzmann statistics and random number generators are utilized to determine whether each new configuration generated is "accepted" or "rejected". The Monte Carlo procedure is as follows: a new configuration is generated by randomly moving atoms or residues and then calculating the energy of the new configuration using a potential energy function. If the energy of the new configuration is lower than the previous, the new configuration is accepted. If the energy is higher than the previous then the 'Boltzmann Factor' of the energy difference is calculated using the following equation:

$$Boltzmann\ Factor\ =\ e^{-(\nu_{new}(r^N)-\nu_{old}(r^N)/k_BT} \qquad \textbf{Equation 1.5}$$

where $v(r^N)$ is the energy of the system calculated by the potential energy function, $k_B$ is the Boltzmann Constant, and T is the temperature. The calculated Boltzmann factor is then compared to a random number generated between 0 and 1 and if it is higher, the configuration is accepted, if it is lower, the configuration is rejected. In the case of rejections, the original configuration is then used again to generate a new configuration. The use of Boltzmann statistics with a given potential energy function ensures that configurations with lower energies are generated more frequently than higher energy configurations. The result is an ensemble of configurations for which desired property averages or positional averages can be calculated.

There are several key differences between dynamics methods and Monte Carlo methods which should be pointed out. First, as mentioned above, dynamics are a deterministic method which can provide information about a system which is time-dependent. Because Monte Carlo methods are random, there is no temporal relationship between the configurations generated. Second, MD methods include a kinetic energy component when calculating the total system energy, where in MC methods; the total system energy is calculated from a potential energy function. Finally, MC methods have the ability to sample higher energy configurations which may play a role in structure and function, but are harder to reach using MD methods. Because Monte Carlo methods were not employed in the studies discussed herein, we will limit our discussion below to MD approaches and practical considerations.

### 1.3.2 Molecular Force Fields and Parameterization

Molecular force fields, or molecular mechanics, are the backbone of molecular dynamics simulations. Force fields, which were mentioned above in the docking-based virtual screening section, are energy functions which are used to calculate energies of molecular systems based only upon the atomic positions and do not take into account electronic effects like quantum mechanics and semi-empirical methods. Because of this, force fields cannot be used to describe molecular properties that depend upon electron distribution, such as chemical reactions. Force fields describe the energy of a system using a series of energy terms that can be generally classified as internal (bonded) terms and external (non-bonded) terms. The energy of the system is the sum of all the internal and external terms calculated for all atoms, bonds, and interactions in a given system. Typical bonded energy terms in a simple force field include bond length (or bond stretching), angle bending, and torsional rotation, while the non-bonded terms include energy terms to describe electrostatic and van der Waals interactions, usually a Coulomb potential term and a Lennard-Jones potential term, respectively. The functional form for such a simple force field is given by the following equation:

$$v(r^N) = \sum_{bonds} \left(\frac{k_i}{2}(l_i - l_{i,0})^2\right) + \sum_{angles} \left(\frac{k_i}{2}(\varphi_i - \varphi_{i,0})^2\right) + \sum_{torsions} \left(\frac{V_n}{2}(1 + \cos(n\omega - \gamma))\right) + \sum_{i=1}^{N} \sum_{j=i=1}^{N} \left(4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\right)$$

**Equation 1.6**

where the bonding term and the angle term are harmonic potentials that increase in energy as the bond length, $l_i$, and angle length, $\varphi_i$, vary from their reference values, $l_{i,0}$ and $\varphi_{i,0}$. The torsional term describes how the energy changes as the torsion angle changes; $\omega$ is the measured torsion angle, $n$ is the multiplicity (number of minimum energy points as the torsion angle rotates through 360°) and $\gamma$ is the phase factor which determines where the torsion angle passes through its minimum value. The force constants, $k_i$ and $V_n$ in these terms are specially derived (parameterized) so that the force field is able to reproduce known experimental energy or property values when it is being trained or designed. The last term is the non-bonded energy term which includes the 12-6 Lennard-Jones potential for calculating van der Waals interactions and a Coulomb potential for calculating electrostatic interactions.

Equation 1.6 is an example of a simplified force field with the minimum necessary terms to calculate the energy of a system. As mentioned in the scoring section above force fields can include additional terms to describe special types of energetic contributions such as aromatic or pi stacking, H-bonding, solvation/desolvation, and rotatable bond restrictions. Additionally, some advanced force fields can contain 'cross terms' which account for coupling between different energy components, such as bond stretching with angle bending, and stretching and bending with torsional changes. Force fields can be categorized as either Class 1, 2, or 3 depending on the energy terms they incorporate and the complexity of those terms. Simple force fields which contain only harmonic terms and do not contain cross terms (such as the example shown in equation 1.6) are called Class 1 force fields. Class 2 force fields may include anharmonic terms (Morse potentials or cubic and quartic energy term expansions) as well as cross terms while Class 3 force fields take into account chemical effects such as hyperconjugation, polarization, and electronegativity. Table 1.4 lists examples of several commonly used force fields in small molecule and biomolecular modeling. Obviously, as the complexity of the force field increases, so does the time required to complete energy calculations for a system. For the purposes of molecular simulations, Class 1 force fields are typically utilized due their speed when compared to Class 2 and Class 3 force fields. Class 2 and Class 3 force fields are generally used only for performing calculations on small molecules.

**Table 1.4. Types and Examples of Commonly Used Molecular Mechanics Force Fields**

| Type of Force Field | Representative Examples |
|---|---|
| Class 1 (Classical) | Amber,[109,110] CHARMM,[111] OPLS,[112,113] GROMOS,[114] |
| Class 2 (2nd Generation) | CFF,[115] MMFF,[108] MM2/MM3[116-118] |
| Class 3 (3rd Generation) | MM4[119,120] |

An important part of force field modeling is the assignment of the atom type. In quantum mechanical calculations, the atomic number, spin multiplicity, and overall charge of the nuclei present must be provided as input in order to obtain meaningful results. Although force fields don't require electronic information on the system under analysis, some information is still necessary on the types and number of atoms present. Atom types can provide the force fields with information regarding element, hybridization, ionization, and valence. Atom types can be very general or very specific, depending on the nature and purpose of the force field. Table 1.5 gives some examples of atom types with their descriptions from a general force field, the Tripos FF, and the Amber7 FF. The table lists all carbon, nitrogen, and oxygen atom types for the Tripos FF and only carbon atom types for Amber FF. It is readily apparent that the Tripos FF is much less specific than the Amber FF with respect to the number of atom types necessary to describe a system. This is because the Tripos FF is considered a general purpose FF that can be used to describe a wide variety of molecular types, usually small molecules, while the Amber FF has been designed and parameterized to specifically deal with large biopolymers composed of amino acids or nucleic acids (i.e. proteins and DNA).

Atom types play a key role in parameterization of force fields as each force field parameter, as in the force constants described above, is expressed in terms of atom types. For example, in the example given above, there would be reference bond lengths, angles and torsions with corresponding force constants for each combination or set of atom types, two atom types for bonds, three for angles, and 4 for torsions. Parameterization is the process of developing the reference values and force constants for a given force field. This can be a time consuming process, but it is very important as the overall performance of the force field is dependent on the quality of its parameters. Parameterization can be considered a two step process, the first is to identify and define the reference values for each atom type defined in the force field and the second is the assignment of the force constants to be used. Reference values are usually obtained in one of two ways, from experimental data or from quantum mechanical calculations.

**Table 1.5. Example Atom Types from Tripos and Amber Force Fields**

| Tripos FF | Description | Amber7 FF | Description |
|---|---|---|---|
| C.3 | sp3 carbon | C | any carbonyl sp2 C |
| C.2 | sp2 carbon | C* | sp2 aromatic C in 5-membered ring |
| C.1 | sp carbon | CA | any aromatic sp2 C |
| C.ar | aromatic carbon | CB | sp2 aromatic C at junction between 5- and 6-memberd rings |
| N.3 | sp3 nitrogen | CC | sp2 aromatic C in 5-membered ring with 1 substituent and next to an N |
| N.2 | sp2 nitrogen | CD | sp2 C atom in C=CD-CD=C |
| N.1 | sp nitrogen | CK | sp2 aromatic C in 5-membered ring between N and N-R |
| N.ar | aromatic nitrogen | CM | any sp2 C, double-bonded |
| N.pl3 | trigonal planar nitrogen | CN | sp2 aromatic junction C between 5- and 6-membered rings, bonded to CH and NH |
| N.am | amide nitrogen | CQ | sp2 C in 6-membered ring lone pair Ns |
| N.4 | sp3 positively charged nitrogen | CR | sp2 aromatic C in 5-membered ring between 2 Ns |
| O.3 | sp3 oxygen | CT | any sp3 C |
| O.2 | sp2 oxygen | CV | sp2 aromatic C in 5-membered ring between 2 a C and lone pair N |
| O.co2 | oxygen in carboxylate | CW | sp2 aromatic C in 5-membered ring bonded to a C and an N-H |
| O.spc | oxygen in SPC water model | CY | nitrile C |
| O.t3p | Oxygen in TIP3P water model | CZ | sp C |

Once the reference values are obtained, the force constants are usually developed by "fitting" the force field to experimental data, which can be thermodynamic properties of a system, known binding energies, or other properties obtained from quantum mechanical calculations. This involves stepwise modification of the force field parameters give progressively better fits to the data being used. Thus, parameterization is an iterative process. Fortunately, most of the available commercial or academic force fields have been well parameterized for use against the systems they were developed. However, in some cases the molecular modeler will have to develop and add parameters for compounds, atom types, bond types, etc. that are not explicitly described in the force field being used.

### 1.3.3 Molecular Dynamics Approaches and Practical Considerations

Setting up and running a molecular dynamics simulation is a complicated process which requires many considerations, such as the initial configuration of the system being studied, choice of force field and dynamics integration method, time length of the simulation and time steps, type of ensemble and energy calculations, boundary conditions, and solvation. Each consideration can influence the outcome of the simulation as well as the computational expense and time requirements.

The initial configuration of the system is usually obtained from experimental data, theoretical models, or a combination of both. For example, for a protein simulation, the structure of the protein may have been obtained from x-ray crystallography, NMR, or homology modeling. Atom types for the force field being used must be defined and parameters developed if necessary. Partial atomic charges are loaded using one of the methods described above. Finally, the systems are frequently minimized prior to running dynamics to eliminate high energy interactions such as steric clash.

The force field (see above discussion) and the integration method are chosen based upon the nature of the system, i.e. small molecule, DNA, protein, etc., and the information desired from the simulation. Another consideration is how well the dynamics program to be used can be parallelized. Parallelization is very important as large biomolecular simulations must be run across multiple processors or on "clusters" in order to be completed within a reasonable amount of time. Molecular Dynamics packages available for commercial or academic use frequently incorporate their own force field which has already been parameterized by the developers. Some commonly used dynamics packages that include their own force fields are Amber, CHARMM, and GROMACS.[72,73,75] There are also a number of packages available for academic uses that utilize other developed force fields. For example, LAMPPS[121] is compatible with the CHARMM, AMBER, OPLS, and GROMOS force fields; NAMD[74] can be used with CHARMM, AMBER, and OPLS; and GROMACS[75] can be used with its native force field GROMOS, CHARMM, or AMBER. There are a variety of integration methods currently employed by dynamics software packages, including the Verlet algorithm,[122] the 'leap-frog' algorithm,[123] the velocity Verlet method,[124] and Beeman's algorithm.[124] Factors that

must be considered when choosing an integration method include computational effort required, length of time steps required, energy conservation, and the ability of the methods to deal with the ensemble method being used. The most widely employed integration methods employed today are Verlet and velocity Verlet methods. The Amber package employed in the Molecular Dynamics studies described in the next chapter uses the velocity Verlet integration method by default.

Once the initial configuration of the system has been defined and the force field and integration method (software package) selected, decisions must be made as to the length of time and the time steps that will be required for the simulation. The length of time will be determined by the nature of the system being studied, the process being studied, and the computational resources available to the modeler. Currently, the time length limitation for dynamics simulation is on the order of tens to hundreds of nanoseconds, although microsecond simulations for smaller systems have been reported. For example, protein folding, which occurs on a millisecond time scale, is not currently observable using molecular dynamics methods, but small scale loop movements and ligand binding can theoretically be observed on the nanosecond or low microsecond time scales. The calculation time steps are another key consideration and will depend on the integration method being used, the system studied, and the computational resources available. Obviously, the smaller the time step chosen, the more computational expensive will be the simulation and the resulting time required to complete the simulation will increase. A standard recommendation is that the time step chosen should be one-tenth the time of the shortest motion being studied. In biomolecular systems this is usually the C-H bond vibration which occurs on a 10fs time scale, thus 1fs time steps would typically be chosen. If C-H bonds are held constrained during the simulation using a method known as the SHAKE algorithm, then this time step can be doubled to 2fs.[125]

The next consideration is the type of ensemble to be studied and the types of energy calculations that will be used. Molecular dynamics are traditionally performed using the NVE or microcanonical ensemble, which holds constant the number of particles (N), the volume (V) and the energy (E). Monte Carlo methods traditionally utilize the NVT or canonical ensemble (constant N, V, and temperature, T). When studying biomolecular systems, however, it is more practical to use the NTP, or isothermal-isobaric ensemble, which holds constant the number of particles (N), the temperature (T), and the pressure (P). This simulates physiological conditions more closely than the other types of ensembles. Probably the most time consuming part of a molecular dynamics simulation is the calculation of long range interactions and there are a variety of methods for handling this. The use of distance cutoffs for energy calculations is one popular way to address this problem. Cutoffs present a problem with certain types long-range interactions, such as charge-charge interactions which can still significantly contribute to the energy of the system beyond the standard cutoffs used in most dynamics simulations. Special methods have been developed to address this problem, including the Ewald summation, the reaction field method, and the cell multiple

method. The Ewald summation method is probably the most popular, and a version known as Particle-Mesh Ewald (PME) is currently deployed in the Amber simulations package.[126]

Finally, boundary conditions and choice of solvation methods must be decided upon. Because interactions at the boundaries of the system being studied (i.e. vacuum, wall, etc.) can influence the energy calculations, the boundaries must be defined or taken into account in some manner. For biomolecular simulations, the most common way to do this is to employ periodic boundary conditions. Periodic boundaries involve placing the system in a cell, typically a cubic box or other geometric shape, and then surrounding the cell with mirror cells containing replicas of the system (26 cells for a cubic box). The interactions energies can then be calculated across cell boundaries overcoming the boundary effect and essentially enabling the simulation of a much larger system. If a particle leaves one side of the cell, it subsequently enters from the other side; this keeps the number of particles in the system being studied constant. One caveat that must be mentioned here is that the cell size chosen must be large enough so that the actual biomolecule being studied does not "see" itself and affect its own energy calculations. Usually, it is desirable only for solvent molecules to cross the periodic boundary.

This brings us to the final consideration, the choice of solvation method. There are currently three different ways to take into account solvation: the first is to simulate the system *in vacuo* using only a distance dependent dielectric screening term in the force field to simulate the solvent screening effects on electrostatic charge calculations. This method is the least rigorous, eliminates the need for periodic boundary conditions, and is the fastest in terms of computational expense; however it is also the least reliable and should be reserved only for simulations where solvent effects are not expected to play a key role. The second method is to model solvation in a dynamics simulation is known as 'implicit solvation', or continuum solvation. This method uses special energy terms in the force field to represent the solvent as a continuous medium. Two commonly used algorithms are used to approximate the solvent electrostatic effects: the Poisson-Boltzmann equation, and the Generalized Born model, which is a linear approximation of the Poisson-Boltzmann equation that is less computationally expensive. Both of these equations are often combined with a hydrophobic solvent accessible surface area (SA) term. Implicit solvation models, while more reliable than in simple dielectric terms, still have limitations. Entropic effects are not accounted for in these models, which can be a major factor in loop movements, ligand binding, and protein folding. The effect of solvent viscosity on the motion of solutes is also not accounted for when using implicit models, although in some cases this can be desirable. Finally, although H-bonding can be generally accounted for with implicit solvation algorithms, the directionality of H-bonds cannot. The final solvation method is known as explicit solvation. In this method the solvent molecules are explicitly treated by surrounding the solute or biomolecule by solvent molecules. This method is the most accurate but also the most computationally expensive as all energy calculations must now include the many solvent molecules,

typically on the order of 50,000 or more, needed to solvate the biomolecule. For biomolecular simulations, there are several water models that have been designed for use, the most commonly used is the TIP3P water model, a 3-site model where the water is represented by a molecule with 3 interaction sites and a rigid shape.[127] 4, 5 and 6 site models have been developed but they increase the computational expense of the simulation and are rarely used except for simulations modeling water dynamics.

Once the molecular dynamics methods have been determined and the system has been set up, the simulation can be run. A typical dynamics simulation of a biomolecular system under explicit solvation is a multi-step process. An initial solvent minimization is required, where the solvent is minimized while the solute is held under constraint. This is followed by a solvent dynamics step, where the solvent (and any counter ions added to balance the solute charge) are allowed to equilibrate; typically 10 to 100ps are sufficient. The next step would be allowing the entire system to minimize while slowly loosening the constraints on the solute, or biomolecule. This is followed by the dynamics simulation itself which occurs in two phases, an equilibrium phase and a production phase. The equilibrium phase brings the system to equilibrium from the starting configuration, often while raising the temperature slowly to the desired simulation temperature. Equilibration is reached when the calculated average temperature, pressure, and energies have stabilized. Finally, the production phase of the simulation can begin, where the system is allowed to fully evolve for the desired time period. Typically only data obtained from the production phase is used to calculate the desired properties.

## 1.4 Contemporary Structure-Based Drug Design

Over the last decade, a number of new drug design techniques have emerged that are gaining wide acceptance in industry and academia. This section will introduce several contemporary design techniques that can be incorporated into a structure-based drug design program, discuss methods involved with these techniques and examples of their successful application. The focus of this section will be fragment-based drug design with a brief introduction to click chemistry, tethering and dynamic combinatorial diversity.

### 1.4.1 Principles of Fragment-Based Drug Design

In 1997, Lipinski et al. proposed the “Rule of Fives” for drug-likeness, solubility and oral bioavailability.[23] The model proposed that an ideal oral drug candidate should have a molecular weight of no more than 500 Daltons, no more than 5 hydrogen bond donors and 10 hydrogen bond acceptors, and a ClogP no greater than 5. This model was readily adopted by both the pharmaceutical industry and academia and is commonly used to filter corporate libraries and large compound collections prior to high-throughput screening. In fact, many drug companies have fashioned their corporate

libraries to be in compliance with the Rule of Fives. However, a study published in 1999 by Teague, et al. which examined the lead compounds for a large number of commercially available drugs demonstrated that the molecular weight, logP, rotatable bonds, and hydrogen bond donor and acceptor counts were significantly lower than the final marketed compounds.[128] They concluded that the lead optimization process almost always leads to more complex compounds, and advanced the concept of lead-likeness versus drug-likeness. The authors went on to propose that screening programs should focus on lead-like or fragment compounds rather than drug-like compounds.

Since that time, many other studies have been performed investigating the concept of fragment-based screening, and many groups have reported the discovery of novel compounds with low nanomolar potency utilizing this methodology.[129] There are now definite criteria for defining lead-like or fragment compounds, the most commonly used being Congreve's "Rule of Three" which states that most fragment hits have a molecular weight of ≤ 300, ≤ 3 hydrogen bond acceptors, ≤ 3 hydrogen bond donors, and have a ClogP of ≤3. Additionally, the rotatable bond count should be ≤ 3 and the polar surface area should be ≤ 60 Å.[128,130]

There are several advantages to screening fragments over drug-like compounds, the most notable being the likelihood that the lead optimization process will result in a drug-like compound which has a greater chance of having good oral bioavailability and favorable ADME properties. Compare this with the drug-like screening process, where the optimization of a drug-like hit, which again is most likely to increase molecular size and complexity, could very possibly result in a compound falling outside the desirable range of physicochemical properties for an oral drug candidate. Additionally, because fragments-based methods have the potential to sample higher chemical diversity, a much smaller number of compounds are generally needed for fragment-based screening, usually on the order of hundreds to a few thousand. This concept is best explained using an example. Two fragment libraries each containing 1000 fragments would contain 1,000,000 compounds when combined using a single linker. Screening this number of compounds would be a significant undertaking. However, if one were to test the fragments first, then take the 5 most active from each set and combine them in a similar manner, the result would be only 2025 compounds that required testing (1000 fragments + 1000 fragments +25 linked compounds), a significantly easier undertaking, while still covering the same chemical space. The advantage of fragment screening is obvious when compared to a standard high-throughput screen that would involve hundreds of thousands to millions of compounds.

Another advantage is that fragment-based screening can lead to higher hit rates. This is because compounds of lower complexity have a greater chance of matching the target receptor site.[131] As the complexity of the compounds being screened increases, the probability of binding (hit rate) decreases. Finally, dealing with fragments rather than larger, drug-like compounds, is advantageous from a technical perspective in that data management, compound acquisition, and synthesis are all simplified.

### 1.4.2 Fragment-Based Drug Design Methods

As previously discussed, during the lead optimization process, a fragment hit of low potency can be developed into a drug-like compound of very high potency. There are several optimization techniques which have been used with success that deserve special mention: fragment evolution, fragment linking, and fragment self-assembly. In fragment evolution, additional functional groups are added to the fragment hit to optimize binding and increase potency. This process is generally guided by X-ray or NMR structural information. One requirement to using this optimization procedure is that the original hit fragment must act as an 'anchor' and not alter its binding position during the evolution process. This method is most useful for smaller active sites that can afford multiple fragment binding sites.

If the targets active site can accommodate multiple fragment binding sites, then the fragment linking optimization method can be employed. Fragment linking involves the addition of a linking group (which may or may not form receptor site binding interactions of its own) to join two fragment hits that bind into two separate sites on the target receptor. This linking frequently results in compounds whose potency is much greater than that of the two starting fragments, mostly due to entropic considerations. The key point here is that the expected free energy of binding of the linked molecule is greater than the sum of the binding energy of the two individual fragments. This is because with the two separate fragments, there are two entropic penalties to binding, whereas only one with the linked compound. One important consideration when using the fragment linking method is that the two fragments must remain in their original binding positions after being linked. This factor can determine the choice of linking group to be used.

Fragment self-assembly involves the use of chemically reactive fragments that are able to bind into the active site and react with each other, forming a larger inhibitor. The active site acts as a template for the reactive compounds, aligning them for their reaction and filtering out fragments not able to match the active site characteristics.[132] This is an example of "click chemistry" as it applies to fragment-based design.

### 1.4.3 Fragment Activity and Binding Analysis

Fragment-based screening is not without its disadvantages. Because of the lower molecular weight and complexity of the fragment compounds, they are expected to be less potent than a drug-like compound. This means that specialized screening methods need to be employed to identify hit compounds. Several methods have been used with success, including high concentration screening,[133] X-ray crystallographic screening,[134] NMR screening[135], affinity detection by mass spectrometry[136], surface plasmon resonance[137], and isothermal titration calorimetry (ITC).[138] It should be noted that although the fragment hits typically show a much lower potency, often high micromolar to low millimolar, in terms of binding efficiency (binding affinity normalized by

molecular weight or heavy atom count), they are often on par with or exceed the efficiency of drug like compounds.[139] Binding efficiency is a key concept of fragment-based drug design.

Typical high-throughput screening experiments assay the compounds being tested at 10 μM concentration. In fragment-based drug design, because of the lower binding affinity of the fragment 'hit' compounds, a 10 μM concentration is not sufficient to detect activity. Concentrations up to low millimolar must be used to detect active fragments; typical high-concentration fragment screens will use 250 or 500 μM. This presents special problems associated with this high concentration screening. At higher concentrations, solubility of the screening compounds can become an issue and compounds can precipitate out of the screening solution which can interfere with assays and activity detection. Therefore, if high-concentration screening is to be used for activity detection, it is advisable to build the fragment screening library using only very soluble compounds.

With the advent of high-throughput crystallography techniques, x-ray crystallography has evolved into a screening technique that can identify fragment binding. Fragment compounds are typically screened by soaking cocktails of 4 to 10 compounds, which have been selected for optimum diversity, into protein crystals. There are several advantages to screening using this method. First, the binding mode of the fragment can be directly visualized, which facilitates the subsequent lead optimization process. Also, multiple binding fragment binding sites can visualized, which can facilitate the fragment linking approach to lead optimization. And finally, unlike traditional screening methods, high fragment concentrations are not normally necessary when using x-ray crystallography methods. This method has gained considerable acceptance with drug companies in recent years and several have developed crystallography platforms for use specifically with fragment screening, including Astex' *Pyramid*,[140] Stuctural GenomiX' *FAST*, Plexxikon's *Scaffold-Based Drug Design*,[141] and Abbott's *CrystaLEAD* process.[134]

Nuclear magnetic resonance (NMR) is another sensitive method that can be used to detect fragment binding. There are two general methods that can be employed to detect fragment binding: detection by receptor (protein) resonances and detection by ligand (fragment) resonances. When detection is done by observing receptor resonances, and initial map or "fingerprint" of the receptor amide or methyl protons is obtained in a non-bound state which can then be compared to resonances obtained with cocktails of fragments compounds present. Chemical shifts of $^{1}H$-$^{15}N$ or $^{1}H$-$^{13}C$ resonances in the active site can indicate bound ligands. It is even possible to localize the binding site if sequence-specific resonance assignments are available. Ligand based methods take advantage of the differences in ligand resonances between bound and unbound states and typically use one of two methods: Saturation Transfer Difference (STD)[142] and WaterLOGSY.[143] Receptor and ligand based NMR detection methods each have their advantages and disadvantages. A key advantage to using the

receptor-based method over the ligand based method is that high affinity compounds can be detected by analyzing the resonance peaks obtained. Also, because the protein has been assigned, it is possible to identify binding to target and non-target sites on the protein. Disadvantages to receptor-based NMR methods include molecular weight limitations for the proteins (>30 kDa is typically beyond the practical range for protein sequence assignments), requirements for large amounts (milligrams quantities) of protein, and long sample stability requirements. Finally, a key disadvantage of ligand-based detection methods is that tight-binding ligands can show as false negatives because they do not disassociate from the receptor frequently enough to distinguish between bound and unbound ligand resonances. SAR by NMR[144] is a specialized fragment linking technique that allows for the design of high affinity ligands by linking lower affinity fragments that have been detected by 2D NMR methods. The key to the SAR by NMR method is that two separate binding sites and ligands have to be identified and linked using fragment linking methods.

The last three methods of binding and activity analysis are mass spectrometry (MS), Surface Plasmon Resonance (SPR), and isothermal titration calorimetry (ITC). Electrospray ionization is typically used in MS techniques to ionize protein/ligand complexes. Mass identification can then be used to identify fragment binding, even from mixtures of fragments. This can be a sensitive detection method, but relies on the ability of the protein/ligand complex to remain together in a gas phase ionized state. Disadvantages include the requirements for relatively large quantities of protein and an unclear understanding of the effect ligand binding forces between ligand/receptor on going from solution to gas phase. In the SPR detection method, the protein is typically immobilized onto the surface of a solid support after which screening compounds are introduced. Binding is detected by analyzing changes in the refractive index at the surface caused by co-localization of the ligand and protein.[145] This method has advantages in that it is possible to measure kinetic binding data and there are no affinity limitations. However, because the protein is immobilized the measurements do not take place in solution. Also, a method of immobilizing the protein in its active state must be utilized. Finally, isothermal titration calorimetry has shown utility in the identification of low-affinity compounds in recent fragment-based design studies.[138] ITC has been widely used to measure the thermodynamic properties of ligand binding by measuring heats of association for receptor-ligand complexation at a given temperature as one component is titrated into the other for complexes involving high-affinity ligands. Using this method the enthalpy of binding (ΔH°), Gibbs free energy of binding (ΔG°), and the disassociation constant ($K_d$) can be determined. From these values one can determine the entropy of binding (ΔS°), by using the following equation:

$$T\Delta S^\circ = \Delta H^\circ - \Delta G^\circ \qquad \textbf{Equation 1.7}$$

One major disadvantage of ITC measurements is that they have been reported to be reliable for low-affinity systems, such as fragment-based studies. However, it has recently been suggested that with ITC measurements can accurately predict $K_d$ values

for low-affinity ligands (~ mM) using improved sensitivity measurements and carefully designed guidelines.[146]

## 1.5 The Design of Antimicrobial Agents: Special Challenges to Computer-Aided Drug Design

The design of pharmaceutical agents with activity against bacterial targets presents some unique challenges and opportunities that will be discussed in this section. Because bacteria are prokaryotic organisms, there are significant differences in these cells when compared to the eukaryotic cells of their mammalian hosts. The metabolic pathways, structural features, and cell components commonly targeted in drug design programs are often unique to bacteria. While this provides an excellent opportunity in terms of selectivity and decreased toxicity, there are also special factors that must be considered, including distribution to the target, bacterial cell penetration, metabolism, elimination, and bacterial resistance. These factors can play a key role in the design of a compound library for screening against bacterial targets.

### 1.5.1 Penetration of Cell Wall

One of the most significant differences between bacterial cells and the cells of their human hosts is the presence of a cell wall. Bacteria can be generally classified by their cell wall dye staining characteristics as either Gram positive or Gram negative. Gram positive bacteria have a simple cell wall located externally to the bacterial cytoplasmic lipid membrane that is primarily composed of a thick layer of peptidoglycan, a series of peptide cross-linked polysaccharide chains. Gram negative bacteria have a more complex cell wall that is composed of a thinner peptidoglycan layer which is covered by a second lipid membrane which contains channels known as ‘porins’. Aside from their functional purpose of maintaining cell stability and structure, the cell wall can present a significant barrier to the penetration of the antibacterial compound in to the cell. Fortunately, the cell wall also presents an attractive antibacterial drug target.

Because of the significant structural differences between Gram positive and Gram negative cell walls, the drug compounds that target Gram positive bacteria are often very different in terms of structure and physical properties from those that target Gram negative bacteria. Gram positive cell walls do not contain the porin channels that are found in Gram negative cells, necessitating the passive diffusion of drug compounds targeting these bacteria across the cell wall. Due to this need for cell wall diffusion, gram positive agents are usually more lipophilic than antibacterial compounds that target Gram negative bacteria. Agents targeting Gram negative bacteria typically enter the cell by crossing through the porins. Because of this these compounds are often more hydrophilic to increase their solubility and facilitate passage through the porin channel.

### 1.5.2. Special Pharmacokinetic Issues to Consider

Figure 1.7 shows the cLogP and molecular weight distribution of the most common antibacterial drug classes. Note the low cLogP of the carbapenems and aminoglycoside antibiotics, two classes commonly used in the treatment of Gram negative infections. This indicates that these classes are very highly water soluble. The penicillin and macrolide antibacterial classes, both commonly used to treat Gram positive infections, are generally distributed into a higher cLogP range, indicating that these compounds are much more lipophilic. Some key general features of antibacterial agents can be seen from Figure 1.7 as well. Unlike drugs for mammalian targets, the antibacterial agents generally have a molecular weights and cLogP values that fall outside of the normally accepted range for "good" oral drug candidates.[23] There are two reasons for this, the first is that many of these classes of drugs have been derived from natural products, which tends to yield compounds with higher molecular weights. The second reason, as has been discussed above, has to do with the unique cell penetration requirements of antibacterial agents.

The trend toward higher molecular weight and decreased lipophilicity seen with several of the antibacterial drug classes has resulted in special pharmacokinetic issues that must be considered. First, oral absorption of the classes with very high MW and low cLogP is significantly decreased, resulting in many agents that can only be given by the intravenous route, such as the carbapenem β-lactams and the aminoglycosides. The route and mechanism of elimination for these compounds is also affected by their molecular weight and lipophilicity. Compounds with high solubility (low cLogP) are primarily eliminated by the kidneys without first being metabolized, while compounds with low solubility (high cLogP) are primarily metabolized prior to elimination. The distribution of these agents to the target tissue is also affected by their high molecular weight and low cLogP. The combination of poor oral absorption and low distribution for several antibacterial drug classes has necessitated the use of large doses, often on the gram scale, in order to the required therapeutic concentrations for efficacy.

### 1.5.3 Screening Library Design for Antimicrobial Targets

The issues discussed in the previous section can strongly influence the design of a screening library to be used against bacterial targets. Special consideration must be give to the nature of the target as well as the classification and cell wall characteristics of the bacteria. As discussed in section 1.2.3, when building a virtual screening library it is often necessary to use filters to focus the screening library so that it contains only compounds that fall within "drug-like" or "lead-like" ranges for molecular weight, lipophilicity, etc. These ranges have been defined by Lipinski, Veber, Congreve, and others for drugs to be delivered by the oral route.[23,24,130] It must be considered, however, that the ranges specified in these studies have predominately been defined by marketed orally available drugs that interact with human targets. As mentioned above, the nature of the bacterial cell and targets that lie within have resulted in average molecular weights

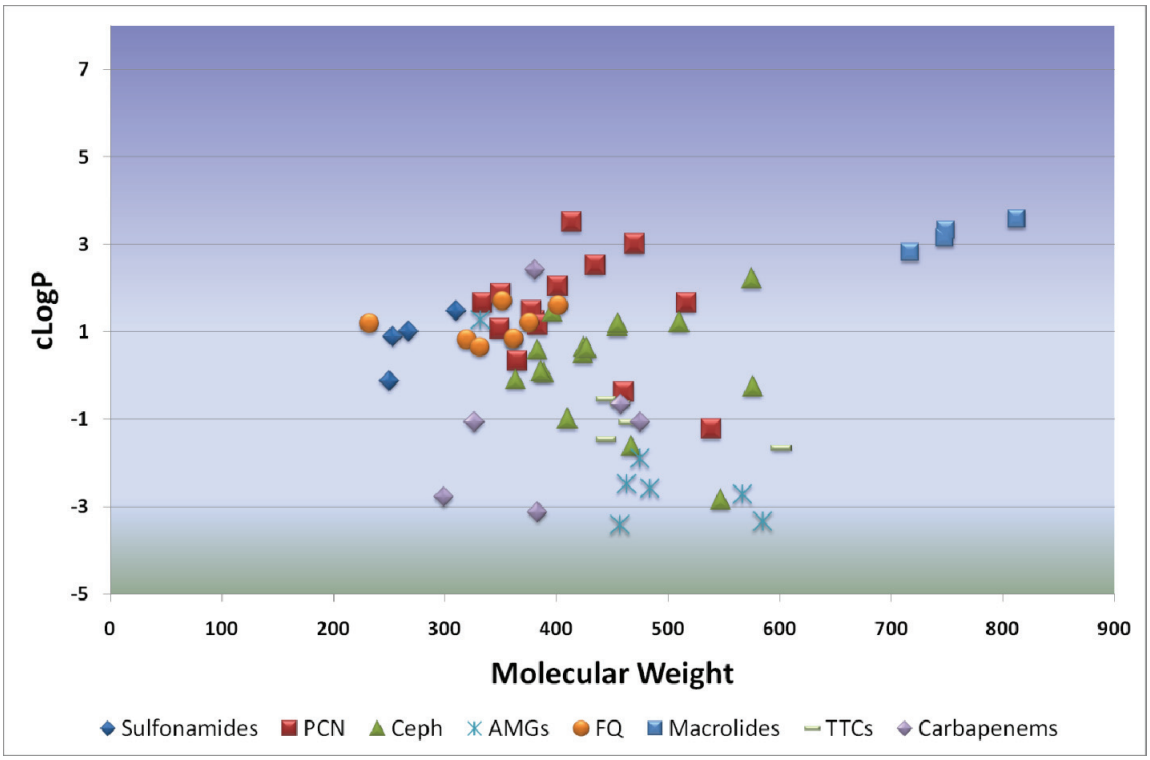



**Figure 1.7. Molecular Weight and cLogP Distribution for Common Antibacterial Drug Classes**

that fall above the “drug-like” range of Lipinski, et al, and cLogP values that fall below these ranges for the drugs that bind these targets. Therefore, when creating a drug-like library for screening against bacterial targets, it is advisable to use higher molecular weight and lower cLogP restrictions. For example, a molecular weight restriction of 650 or 700 daltons rather than 500 is not unreasonable. Similarly, a lead-like library should have slightly higher molecular weight restrictions, on the order of 350 or 400 daltons.

### 1.5.4 Resistance Development

The last issue that will be discussed here is bacterial resistance. It seems that almost as soon as a new class of antibacterials reaches the market, bacteria are isolated that have become resistant. Take for example, the drug linezolid (Zyvox®), which was approved in the U.S. in April, 2000 for the treatment of resistant staphylococcal infections. The first case of clinical linezolid resistance was reported just two years later.[147] This is not a new phenomenon; bacteria have been developing resistance to antibacterials for as long as we have been designing them, as can be seen from the data presented in Table 1.6.

There are several reasons for the rapid emergence of resistance in bacterial organisms, the first is evolutionary. The rapid replication rate of bacteria and the selective pressure applied when treated with antibacterial agents result in the selection of organisms with resistance to these agents. Complicating this is the misuse and overuse of antibiotics and antibacterials, both in the treatment of human infection and use in the environment. Finally, noncompliance on the part of patient for whom antibacterial agents are prescribed also contributes to the rapid emergence of

**Table 1.6. Introduction and Development of Resistance Timeline for Common Antibacterial Drug Classes**

| Antibacterial Class | Mechanism of Action | Introduced | Resistance[a] |
|---|---|---|---|
| Sulfonamides | Inhibit Folate Production | 1935 | 1940 |
| β-lactams (i.e. penicillins) | Inhibit Cell Wall Synthesis | 1942 | 1945 |
| Aminoglycosides | Inhibit Protein Synthesis | 1944 | 1959 |
| Tetracyclines & Related | Inhibit Protein Synthesis | 1948 | 1953 |
| Macrolides & Related | Inhibit Protein Synthesis | 1954 | 1988 |
| Vancomycin | Inhibits Cell Wall Synthesis | 1956 | 1985 |
| Fluoroquinolones | Inhibit DNA Replication | 1985 | 1991 |
| Streptogramins | Inhibit Protein Synthesis | 1999 | 2001 |
| Oxazolidinones | Inhibit Protein Synthesis | 2000 | 2001 |
| Daptomycin | Disrupts Cell Membrane | 2003 | 2004 |

a. Unless otherwise cited, resistance emergence dates are approximate estimates based upon anecdotal reports.

resistance. Bacteria acquire their resistance in one of two ways, either by spontaneous genetic mutation, or by transfer of genetic material from one organism to another.

Bacteria have developed a variety of mechanisms to survive exposure to antibacterial agents. Some common mechanisms of bacterial resistance include: deactivation of the antibacterial agent by enzymatic modification of the compounds structure, decreased permeability of the bacterial cell by altering the cell wall or decreasing porin expression, export of the antibacterial agents by efflux pumps before they can affect their target, alteration of the target's active site such that it maintains activity but no longer affinity for the antibacterial agent, protection of the target by producing biomolecules that interfere with binding, overproduction of the target biomolecule or the natural substrate for the target, and finally utilization of alternate pathways that bypass the inhibited process or pathway. Table 1.7 lists the most common antibiotic drug classes and the most frequent mechanism of bacterial resistance to each class.

A number of mechanisms have been developed that can aid in bypassing the resistance mechanisms mentioned above. In the case of enzymatic inactivation of the antibacterial, compounds can be utilized in conjunction with the antibacterial that inhibit the deactivating enzyme, allowing the antibacterial agent to produce its effect. A classic example of this case is the use of β-lactamase inhibitors with β-lactam antibiotics. The use of two antibacterial agents that inhibit successive steps in a pathway that is being targeted is known as sequential blocking. The best example of this is the use of the antifolate compounds sulfamethoxazole and trimethoprim, the former targeting dihydropteroate synthase and the latter dihydrofolate reductase, sequential steps in the bacterial folate pathway. Efflux pump inhibitors are being investigated for use with antibacterial classes such as the tetracyclines and fluoroquinolones, for which efflux is a major resistance mechanism. Finally, the use of multiple agents that bind to the same target can bypass the altered target resistance mechanism and delay the development of bacterial resistance in some cases.

**Table 1.7. Common Bacterial Resistance Mechanisms Affecting Antibiotic Classes**

| Resistance Mechanism | βL[a] | AG[b] | Mac[c] | Sulf[d] | TCN[e] | FQ[f] | SG[g] | GP[h] |
|---|---|---|---|---|---|---|---|---|
| Enzymatic Inactivation | +++ | +++ | + | - | + | - | - | - |
| Decreased Cell Permeability | + | + | ++ | - | + | + | + | + |
| Efflux | + | + | ++ | - | +++ | + | - | - |
| Altered Target Site | ++ | ++ | +++ | ++ | + | +++ | +++ | +++ |
| Protection of Target | - | - | - | - | ++ | + | - | - |
| Overproduction of Target | - | - | - | ++ | - | - | - | + |
| Bypass of Inhibited Process | - | - | - | + | - | - | - | - |

a. β-lactams
b. Aminoglycosides
c. Macrolides
d. Sulfonamides
e. Tetracyclines
f. Fluoroquinolones
g. Streptogramins
h. Glycopeptides

+++ Most Common
++ Common
+ Less Common

Adapted with permission from Opal, S. M. and Medeirus, A. A. Molecular Mechanisms of Antibiotic Resistance in Bacteria. In *Principles and Practice of Infectious Disease*, 6th ed.; Mandell, G. L.; Bennett, J. C.; Dolin, R., Eds. Elsevier: Philadelphia, 2005; Vol. 1, pp. 253-270.[148]

# CHAPTER 2. STRUCTURAL AND MECHANISTIC STUDIES ON DIHYDROPTEROATE SYNTHASE

## 2.1 Introduction

This chapter and the following two chapters will discuss our efforts in the design of novel molecular agents with activity against the enzyme dihydropteroate synthase. Our work toward this goal followed an approach very similar to that shown in Figure 1.5 in Chapter 1, using a combination of pharmacophore searching and docking using the known crystal structures and known inhibitors of the target enzyme. The target enzyme of these studies is from *Bacillus anthracis*, the causative agent of the disease anthrax, and several crystal structures of the *B. anthracis* DHPS were utilized in this work. This chapter will discuss a series of molecular dynamics simulations that were performed to map the positions of two flexible loops from DHPS, which were unclear in our crystal structures. This was done primarily to build a model that could be used for virtual screening studies, with additional goals of gaining insight into the structure of the transition state and the mechanism of the reaction that DHPS catalyzes. Chapter 3 follows this with a discussion of the docking validation studies which were performed to select the best docking and scoring algorithms for use in virtual screening against DHPS, and finally Chapter 4 discusses the virtual screening studies that were performed and presents the results of those studies.

### 2.1.1 DHPS: New Approaches for an Old Target

The rapid emergence of bacterial drug resistance has led to a decrease in the clinical utility of virtually all marketed antibacterial agents and an increased interest in the design and synthesis of new antibacterial agents with novel targets. An alternative approach to antibacterial drug design is to identify the mechanism of bacterial resistance and utilize this knowledge to develop new inhibitors of established bacterial targets. The sulfonamide class of antibiotics was one of the first classes of fully synthetic compounds successfully used for the treatment of bacterial infections. Sulfonamides act by interrupting the folate biosynthetic pathway in lower organisms by targeting the enzyme dihydropteroate synthase, DHPS. These antibiotics mimic the natural substrate, *p*ABA, and act either by competitive inhibition or by the formation of “dead-end” sulfonamide-pterin products. The key steps of the bacterial folate pathway are shown in Figure 2.1 with the antibacterial inhibition steps highlighted. DHPS catalyzes the addition of *p*ABA to 7,8-dihydropterin-pyrophosphate (DHPP), to form 7,8-dihydropteroate, shown below.

Historically, the sulfonamide antibiotics have been used extensively for a variety of gram-positive and gram-negative bacterial infections. Sulfonamides and combinations with DHFR inhibitors such as co-trimoxazole, a sulfamethoxazole-trimethoprim combo, have been used for the treatment of infections by Neisseria,

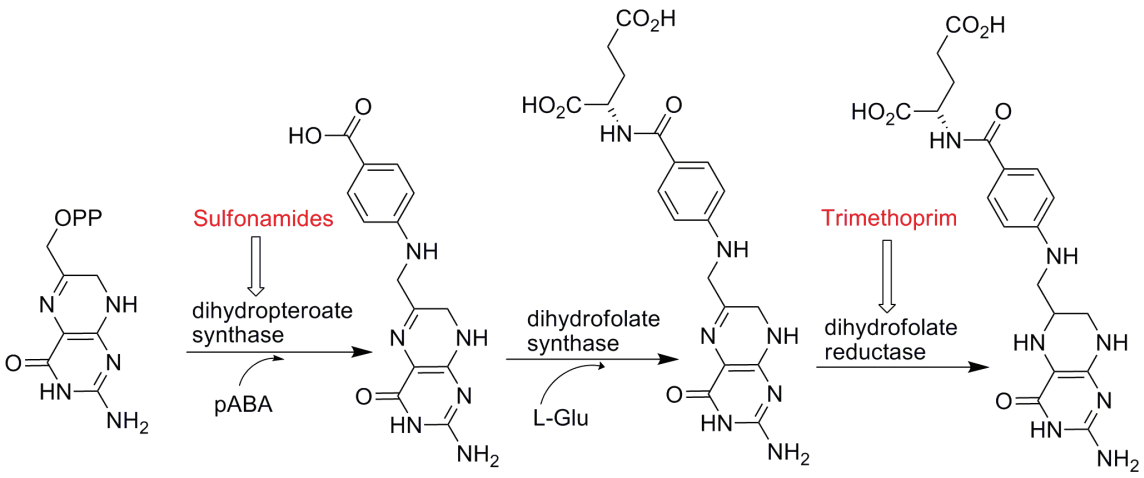



**Figure 2.1. Key Steps in the Folate Biosynthetic Pathway of Prokaryotes**

Streptococci, Staphylococci, Pneumococci, *E. coli*, *Mycobacterium leprae* (leprosy), *Plasmodium falciparum*, and *Pneumocystis jiroveci*. However, drug resistance has emerged as an important factor that severely limits the clinical use of sulfonamide drugs, and resistance mutations in the gene that encodes DHPS, *folP,* have now been characterized in clinical isolates of many pathogenic organisms. This emerging resistance has led to a decrease in the clinical utility of these agents for the treatment of several types of infection, such as upper respiratory tract infections and gastrointestinal infections. Previously considered to be a first-line agent, co-trimoxazole has been relegated to a 2nd or 3rd line option.

Co-trimoxazole is still considered a first line treatment for uncomplicated urinary tract infections and certain types of skin and soft tissue infections, but local resistance patterns often preclude its use. For example, *E. coli*, the most commonly isolated pathogen in urinary tract infections, remains mostly susceptible to co-trimoxazole with a resistance rate of 15-20%, however some urban areas have reported rates as high as 80%.[149] While resistance has certainly caused a dramatic decline in the use of sulfonamide drugs, it should be noted that several emerging pathogens have shown universal susceptibility to co-trimoxazole, lending validity to the further investigation of DHPS and DHFR as drug targets. In fact co-trimoxazole is the recommended agent for treatment of community acquired MRSA, which is rapidly reaching epidemic proportions, and the 7 known clinical isolates of vancomycin resistant S. aureus (VRSA) were all shown to be susceptible as well.[150]

### 2.1.2 A History of Sulfonamide Drug Development

Figure 2.2 shows a timeline of the key developments and discoveries in the history of sulfonamide drug use and development. The discovery of the sulfonamide class of antibacterial agents is credited to Gerhard Domagk, of I. G. Farben Industrie in Germany, who was testing the antibacterial properties of several organic dyes. Domagk noted that the agent Prontosil, shown in Figure 2.3, protected mice against streptococcal infection. Interestingly, Prontosil was only effective when injected directly into mice and had no antibacterial properties when studied *in vitro* against streptococcal species. This was not appreciated until 1935, the same year that Prontosil began to see significant clinical utilization, when Trefouël and coworkers were able to show that Prontosil was metabolized in the body to sulfanilamide (Figure 2.4), and that sulfanilamide was the actual active component. The success of sulfanilamide in the treatment of various streptococcal and staphylococcal infections led to great interest in the development of sulfonamides as antibacterial agents and led to the discovery and utilization of a variety of sulfonamide agents. The actual mechanism of bacterial inhibition was not elucidated until 1940, when Woods and coworkers showed the competitive action of *p*ABA on the effect of the sulfonamides and *p*ABA was subsequently shown to be a key component of folic acid, incorporated during bacterial folate synthesis. This proposed mechanism was not confirmed until 1969, when Richey and Brown were able to purify DHPS and demonstrate the inhibition of DHPS by sulfanilamide.

The picture seemed clear until 1974 when Weisman, Brown, and Bock showed that in some types of bacteria, the sulfonamide agents were actually combined with the pterin substrate to form “dead-end” products, which they theorized went on to inhibit subsequent steps in the folate pathway. However, in 1979 Roland and coworkers showed that a pterin-sulfamethoxazole compound was not able to inhibit DHPS or any other enzyme in the bacterial folate biosynthesis pathway. One possible answer to this conundrum was provided by Swedberg, who proposed that the mechanism of inhibition of bacterial growth by the sulfonamide agents was actually “enzymatic trapping” of the pterin-pyrophosphate substrate in a sulfonamide complex. Swedberg was able to demonstrate a decrease in the effectiveness of these agents when additional pterin-pyrophosphate was added.

### 2.1.3 The DHPS Crystal Structures

The first crystal structure of DHPS (from *E. coli*) was not solved until 1997; a full 36 years after the last sulfonamide agent entered the market. Since that time, crystal structures have been published for six bacterial species (*E. coli*, *S. aureus*, *M. tuberculosis*, *B. anthracis*, *T. thermophilus*, and *S. pneumonia*) and one fungal species (*S. cerevisiae*).[151-157] The DHPS enzyme’s overall structure is a $(\beta/\alpha)_8$ TIM barrel of repeating β/α units which create the classic β barrel composed of eight β strands surrounded by eight α helices. The β strands and α helices are connected by eight flexible loops which fold over the active site in the center of the barrel. Figure 2.5 shows

**1932** Domagk at I.G. Farben observed that Prontosil, an azo dye, protected mice against streptococcal infections. The first patient to be treated was his daughter, Hildegarde Domagk.[158]

**1935** Trefouël, et al demonstrated that Prontosil is converted to sulfanilamide in the body and that sulfanilamide was the active component.[159]

**1936** Colebrook and Kenny demonstrated the efficacy of Prontosil in the treatment of puerperal fever in human beings.[160]

**1939** Domagk was awarded the Nobel Prize. President Roosevelt's son was treated with sulfa drugs, overcoming early reservations.

**1940** The isolation of penicillin reduced interest in sulfa drugs, but the emergence of penicillin resistance renewed interest after WWII.

**1940** Woods, et al demonstrated competition by para-amino benzoic acid (*p*ABA) and the discovery that *p*ABA is part of folic acid pointed to the folate pathway as the target of sulfa drugs.[161,162]

**1961** The last sulfonamide new molecular entity (NME) to be released onto the U.S. market (sulfamethoxazole). The mechanism of bacterial folate biosynthesis was elucidated by Brown, Weisman and Molnar.[163]

**1969** Richey and Brown purified dihydropteroate synthase (DHPS) in the folate pathway, and potent inhibition by sulfanilamide was demonstrated.[164,165]

**1974** Wiesman/Brown and Bock demonstrated the incorporation of sulfonamides into "dead-end' sulfo-pterin products in certain bacteria.[166,167]

**1979** Roland, et al showed that dihydropterin-sulfonamide products do not inhibit DHPS or other folate enzymes. Swedberg theorized that the mechanism of growth inhibition by sulfonamides is enzymatic trapping of pterin-pyrophosphate in a sulfonamide complex.[168,169]

**1997** First x-ray crystal structure of DHPS published, *E. coli*.[151]

**1999** Vinnicombe, et al, demonstrated that the target for sulfonamide inhibition (of *S. pneumoniae*) is the enzyme-DHPP binary complex, rather than the apo form of the enzyme.[170]

**2004** Babaoglu and co-workers solved a crystal structure of *B. anthracis* DHPS with a pterin site inhibitor bound, the basis for the work described here.[154]

**Figure 2.2. History and Key Insights into Sulfonamide Drug Development and Chemotherapy**

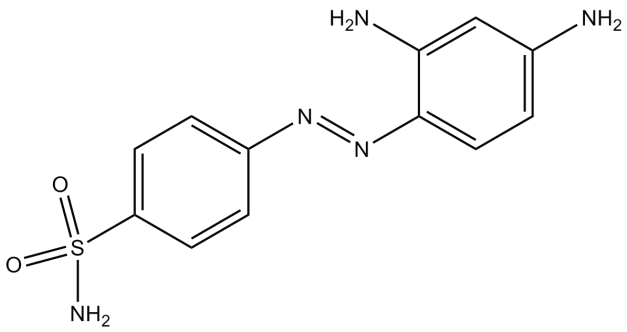



**Figure 2.3. Prontosil**

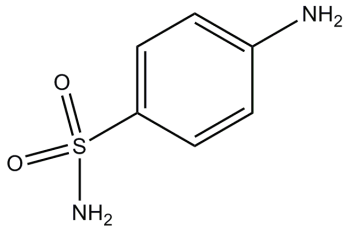



**Figure 2.4. Sulfanilamide**

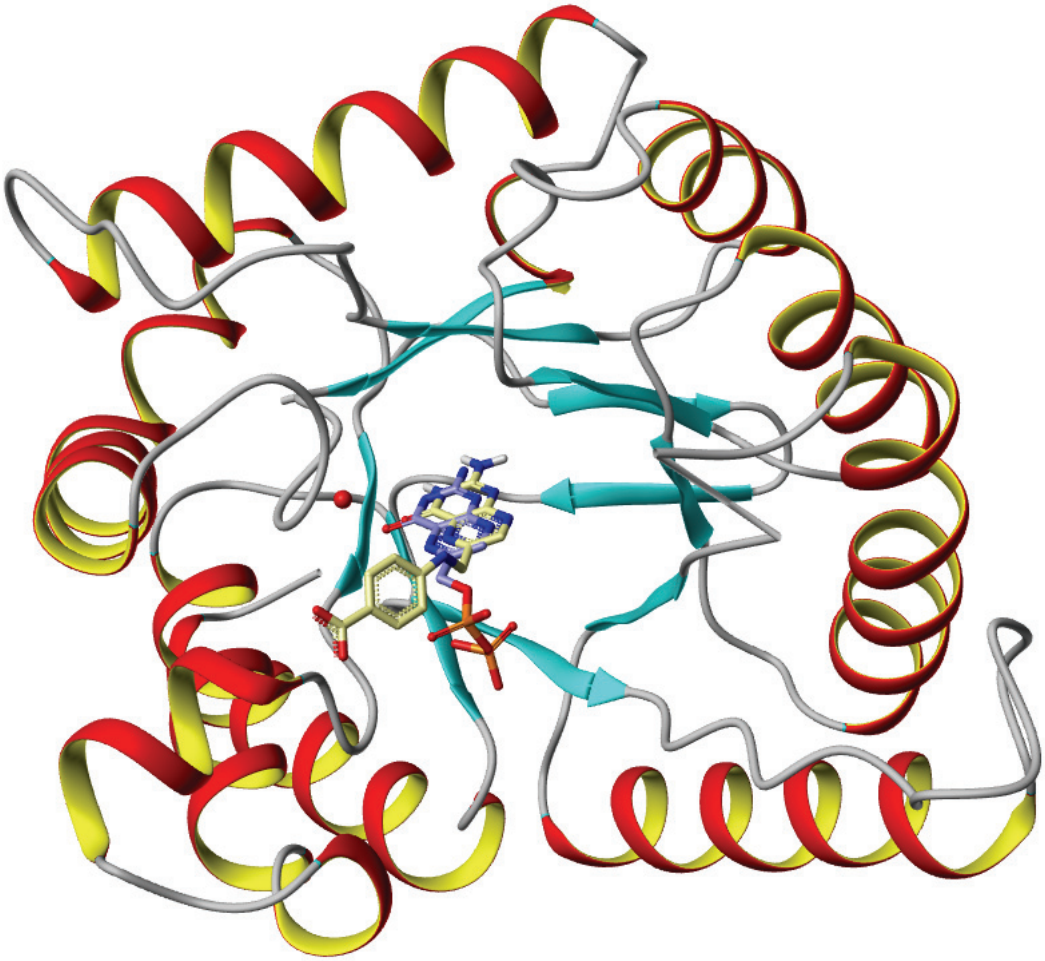


**Figure 2.5. *B. anthracis* DHPS Shown with Product and Substrate Analogs Overlaid**

the crystal structure of *B. anthracis* DHPS with a pteroate product analog and DHPP substrate analog overlaid in the active site. The product analog gives an approximate location for the pterin binding site as well as the *p*ABA binding site, while the substrate analog shows the approximate position of the diphosphate group. The active site of DHPS can be actually divided into 4 distinct subsites: the pterin binding site, the diphosphate binding site, the *p*ABA binding site, and a conserved water binding site, each of which is visible in Figure 2.5. A magnesium cofactor is known to coordinate the diphosphate group and several residues, including His256 and Asn27, and is theorized to play a role in the catalytic mechanism of DHPS. The magnesium ion has not been observed in any of our *B. anthracis* crystal structures to date.

Unfortunately, even with the known structural information, several pieces of the puzzle are still missing and the catalytic mechanism of DHPS remains unclear. The flexible loops that fold over the active site during catalysis enclosing *p*ABA and completing the *p*ABA binding subsite are unresolved or occupy incorrect positions in many of the crystal structures that have been solved. Two key flexible regions in particular, loops 1 and 2, are believed to play a key role in catalysis but are only visible in a few of the structures available and even those positions are uncertain or unreliable. Additionally, as discussed below, the majority of the mutations known to confer resistance to the sulfonamides are found in these two loops. This missing information contributes to our lack of understanding not only of the reaction mechanism, but also the mechanism of sulfonamide resistance. Additionally, the magnesium cofactor mentioned above has only been resolved in 2 of the 17 DHPS crystal structures that have been published to date, and in once case it was replaced by manganese. Table 2.1 gives a listing of the published DHPS crystal structures by species as well as the presence or absence of the key structural features just mentioned for each crystal structure. In some cases, as in our *B. anthracis* structures, it can be seen that although the position of a flexible loop has been solved, it may not be in the transition state or correct binding position, and thus does not contribute to our knowledge of the mechanisms in question.

### 2.1.4 The DHPS Molecular Mechanism: Current Knowledge

Although there is still much uncertainty regarding the exact mechanism of the reaction catalyzed by DHPS, some information is known and some credible theories have been put forth. In their paper presenting the *E. coli* DHPS structure, Achari and coworkers stated their inability to generate a structure with sulfonamide bound by soaking the drug into unliganded DHPS crystals.[151] They were only able to generate a structure when sulfanilamide was soaked along with dihydropterin and pyrophosphate. This result seems to indicate a need for pterin, pyrophosphate, or both to be present in order for the sulfonamide (and presumably *p*ABA) to bind. However, the position of the sulfanilamide in their structure has been called into question by the mechanism proposed by Baca, et al as well as the pteroate product structure by Babaoglu.[154,171] In this latter structure, the position of the pteroate gives an indication of a possible

**Table 2.1. Features of the Known DHPS Crystal Structures**

| Species | PDB Code | Present in Active Site | Loop 1 | Loop2 | $Mg^{2+}$ | Notes |
|---|---|---|---|---|---|---|
| *E. coli* | 1aj0 | Sulfanilamide, $SO_4$, Pterin Analog | Artifactual | Present | Missing | 2.0 Å Resolution |
| | 1aj2 | Pterin-PP Analog, $SO_4$ | Artifactual | Present | Missing | 2.0 Å Resolution |
| | 1ajz | $SO_4$ | Artifactual | Present | Missing | 2.0 Å Apo Structure |
| *S. aureus* | 1ad1 | Nothing | Artifactual | Present | Missing | 2.2 Å Apo Structure |
| | 1ad4 | Pterin-PP Analog | Missing | Present | $Mn^{2+}$ | 2.4 Å Resolution |
| *M. tuberculosis* | 1eye | Pterin-P Analog | Present | Missing | Present | 1.7 Å Resolution |
| *B. anthracis* | 1tws | $SO_4$ | Artifactual | Present[a] | Missing | 2.0 Å Apo Structure |
| | 1tww | Pterin-PP Analog, $SO_4$ | Artifactual | Missing | Missing | 2.5 Å Resolution |
| | 1twz | Pterin-P Analog, $SO_4$ | Artifactual | Missing | Missing | 2.75 Å Resolution |
| | 1tx0 | Pteroate Analog, $SO_4$ | Artifactual | Missing | Missing | 2.15 Å Resolution |
| | 1tx2 | MANIC Inhibitor | Artifactual | Missing | Missing | 1.83 Å Resolution |
| *S. cerevisiae* | 2bmb | Pterin-P Analog | Artifactual | Present | Missing | 2.3 Å Resolution |
| *T. thermophilus* | 2dqw | Nothing | Missing | Missing | Missing | 1.65 Å Apo Structure |
| | 2dza | *p*ABA | Missing | Missing | Missing | 1.90 Å Resolution |
| | 2dzb | Pterin-PP Analog | Missing | Missing | Missing | 1.90 Å Resolution |
| *S. pneumoniae* | 2vef | Nothing | Missing | Missing | Missing | 1.80 Å Apo Structure |
| | 2veg | Pterin-P Analog | Missing | Missing | Missing | 2.4 Å Resolution |

a. In the *B. anthracis* apo structure, loop 2 extends into the pterin binding site replacing the pterin substrate.

transition state in which the *p*ABA group occupies a position significantly different from the position proposed by Achari, et al.

In their paper describing the *S. aureus* DHPS structure, Hampele and coworkers proposed a random, single-displacement reaction mechanism in which the reaction proceeds through a ternary complex of DHPS, DHPP, and *p*ABA. They proposed a random order of addition of substrates based upon their *S. aureus* $V_{max}$ measurements. Vinnecombe and Derrick countered this with their theory that the target for sulfonamide inhibition of DHPS is actually the enzyme-DHPP binary complex.[170] They based this theory on their studies of *S. pneumoniae* and the observation that the *p*ABA substrate binding was absolutely dependent on the presence of pyrophosphate in the active site, which they believed acted as an analog of the DHPP substrate. Additionally, they showed that the sulfonamides displaced *p*ABA in a competitive manner and, interestingly, they also showed that the product of the reaction, dihydropteroate, was also able to bind to the DHPS active site.

In their paper presenting the structure of DHPS from *M. tuberculosis* Baca and coworkers proposed a detailed mechanism and transition state geometry based upon their observations of the Pterin-monophosphate analog in the pterin subsite and the position of the magnesium co-factor. They proposed a trigonal bipyramidal transition state geometry where the C9 carbon of DHPP would develop a partial positive charge which would be stabilized by the electron-rich conjugated pterin ring system. The amino group of *p*ABA would attack the carbon position from the opposite side of the pyrophosphate, as shown in Figure 2.6. The pyrophosphate interacts with a His, Asp, and Asn residue in addition to the stabilizing effect of the magnesium ion, which facilitates the removal of the pyrophosphate during catalysis. Other key binding residues are shown in Figure 2.6. They went on to propose a possible role for a key serine and arginine residue in loop 2 in stabilizing the pyrophosphate during catalysis and facilitating pyrophosphoryl transfer.

Babaoglu and coworkers confirmed the position of the *p*ABA compound proposed by Baca, et al. in their transition state theory with their product analog crystal structure. Additionally, they confirmed the proposed position of the diphosphate group. They proposed that the catalytic mechanism took place in 4 steps. In its unliganded state, a side chain of a key arginine (Arg68 in *B. anthracis*) in loop 2 occupies the pterin binding subsite. In the first step of catalysis the pterin substrate binds and its terminal phosphate group occupies the anion pocket (phosphate subsite). The magnesium cation coordinates the diphosphate group and the $Mg^{2+}$ binding residues mentioned above as loop 2 shifts the arginine residue out of the pterin site. The second step is the formation of the *p*ABA binding subsite by movements of both loop 1 and loop 2. A key part of Babaoglu's theory is the proposed ionic interaction of Arg68 with the *p*ABA substrate, which Babaoglu proposed stabilized the position of *p*ABA. This is in contrast to the arginine position described by Baca, who proposed that the arginine interacted with the terminal phosphate of DHPP during catalysis. The third step would be the

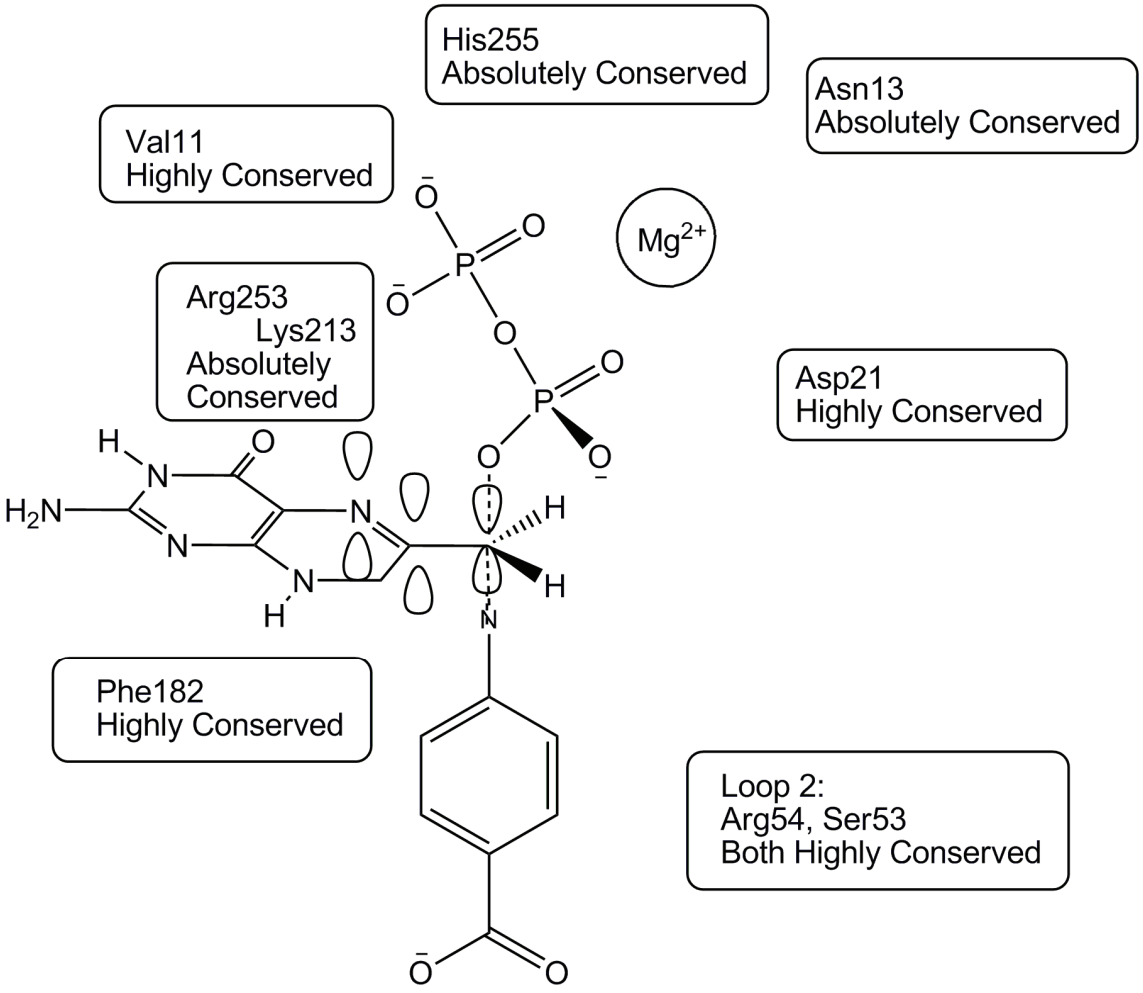



**Figure 2.6. Proposed DHPS Transition State for *M. tuberculosis***

Adapted with permission from Baca, A. M.; Sirawaraporn, R.; Turley, S.; Sirawaraporn, W.; Hol, W. G. Crystal structure of Mycobacterium tuberculosis 7,8-dihydropteroate synthase in complex with pterin monophosphate: new insight into the enzymatic mechanism and sulfa-drug action. *J Mol Biol* **2000,** 302, 1193-212.

nucleophilic attack of the *p*ABA nitrogen on the C9 carbon of DHPP and subsequent loss of a pyrophosphate (facilitated by the magnesium cation), which they proposed took place using an SN2 mechanism, similar to Baca's transition state theory. Finally, the pteroate and pyrophosphate products are expelled from the active site and loop 2 moves back to its position occupying the pterin binding subsite.

There are several questions that have yet to be answered regarding the structure of the transition state and the catalytic mechanism of the reaction. Although key binding residues and several residues believed to be involved in catalysis have been proposed, none of the structures that have been solved to date have shown the positions of both loops 1 and 2 (the location of many of these residues) in their catalytic conformation. The position of *p*ABA during the transition state and even the type of nucleophilic attack are still points of debate as is the function of the arginine residue in loop 2 during catalysis. Does it bind to the terminal phosphate and facilitate its removal or does it bind to the carboxylate of *p*ABA, facilitating the correct alignment of *p*ABA for nucleophilic attack? This remains to be determined. Additionally, the positions and roles of several residues that when mutated confer sulfonamide resistance remain to be determined. This last point is discussed further in the following section.

### 2.1.5 Sulfonamide Resistance Mechanisms

Bacterial resistance to sulfonamide drugs can be caused by a variety of the mechanisms discussed in section 1.5.4, but predominately resistance is caused by chromosomal mutations of the DHPS gene, *folP*, or by the acquisition by the bacteria of plasmids bearing the drug resistant DHPS variants, *sul1*, *sul2* or *sul3*. In the first case, spontaneous mutations of the *folP* gene result in a DHPS enzyme that is no longer capable of binding to sulfonamide agents, but can still bind to the native substrate *p*ABA, albeit usually with decreased efficiency. In the latter case, bacteria have acquired plasmids carrying alternate forms of DHPS which are significantly different from the native enzyme, still capable of binding *p*ABA and catalyzing the reaction with DHPP, but showing markedly decreased affinity for the sulfonamide agents. Notably, with the plasmid variants, the efficiency of *p*ABA binding is not as impaired as in the chromosomally mutated DHPS. Plasmid-borne resistance has only been characterized in Gram-negative enteric bacteria thus far, while chromosomal mutations have been characterized in both Gram-positive and Gram-negative bacteria.[172,173]

Mutations of *folP* conferring resistance that have been characterized in several bacterial species are shown in Table 2.2 along with their corresponding *B. anthracis* positions. The structural positions of these mutations are highlighted in Figure 2.7, which shows *B. anthracis* DHPS with a known pterin inhibitor bound in the active site. It should be noted from Table 2.2 and Figure 2.7 that the mutation sites predominately fall on the flexible loops of the DHPS enzyme with the majority of mutations occurring on loops 1 and 2. Table 2.2 highlights 3 specific mutations that have been observed across several species. In *B. anthracis* these mutations are Phe33Leu(Ile), Thr67Ala(Ile), and

**Table 2.2. Sulfonamide Resistance Mutations Observed from Six Organisms**

| Orgamisms | Mutaton Observed | Corresponding *B. anthracis* Residue | Structure |
|---|---|---|---|
| *E. coli* | Phe28Leu, Ile | Phe33 | Loop1 |
| *N. meningititis* | Phe31Leu | Phe33 | Loop1 |
| *P. carinii* | Phe23Leu | Ser34 | Loop1 |
| *M. leprae* | Thr53Ile, Ala | Thr67 | Loop2 |
| *P. carinii* | Thr55Ala | Thr67 | Loop2 |
| *P. falciparum* | Ser436Ala, Phe | Thr67 | Loop2 |
| *P. falciparum* | Ala437Gly | Arg68 | Loop2 |
| *S. pneumoniae* | Arg58-Pro59 duplication | Arg68-Pro69 | Loop2 |
| *M. leprae* | Pro55Leu | Pro69 | Loop2 |
| *P. carinii* | Pro57Ser | Pro69 | Loop2 |
| *E. coli* | Pro64Ser | Pro69 | Loop2 |
| *S. pneumoniae* | Arg insertion after Gly60 | Gly70 | Loop2 |
| *S. pneumoniae* | Ser61 duplication | Phe71 | Loop2 |
| *P. carinii* | His60Asp | Ala72 | Loop2 |
| *S. pneumoniae* | Ile66-Glu67 duplication | Val74-Ser75 | Loop2 |
| *P. carinii* | Ile111Thr | Ile122 | Loop4 |
| *P. falciparum* | Lys540Glu | Asn147 | Loop5 |
| *N. menignitidis* | Gly194Cys, Ser193-Gly194 duplic. | Gly188 | Loop6 |
| *P. falciparum* | Ala581Gly | Ala190 | Loop6 |
| *P. falciparum* | Ala613Ser, Thr | Gly224 | a7’ |
| *P. carinii* | Val248Gly | Ile246 | a7’ |

Point Mutations Conserved Across Species

Adapted with permission from Baca, et al.[153]

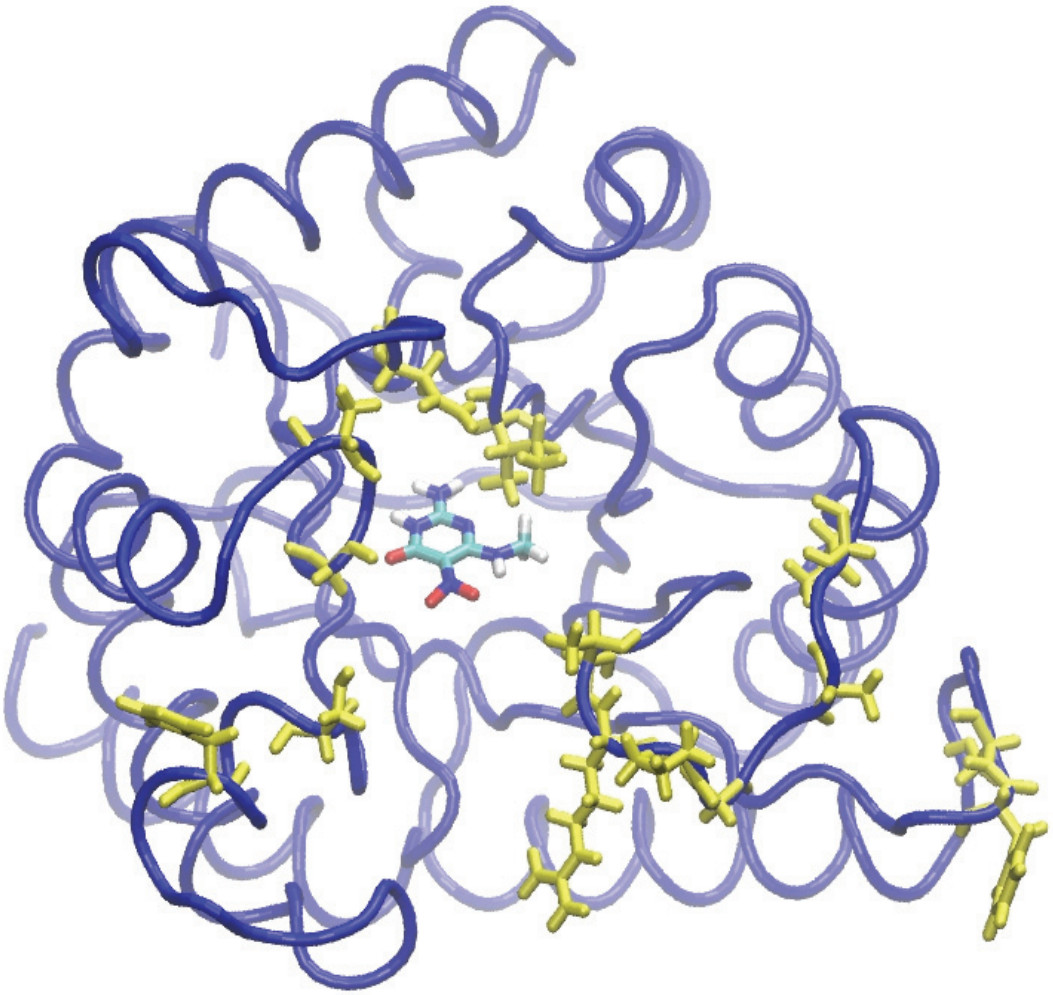


**Figure 2.7. *B. anthracis* Crystal Structure with Known Pterin Site Inhibitor and Key Mutation Residues Shown**

MANIC, a known inhibitor of DHPS, is shown here occupying the pterin binding site in a high resolution X-ray crystal structure. Residues that when mutated are associated with sulfonamide resistance are shown in yellow. It should be noted that Phe33 is pushed far away from the active site due to a crystal lattice interaction with a neighbouring monomer.

Pro69Ser(Leu). Again, all three of these conserved mutations sites fall on either loop 1 or loop2, whose positions remain unresolved or unclear in most of the crystal structures published to date. Thus, the exact mechanism of these mutations in decreasing the binding affinity of the sulfonamide agents has not been determined.

Although the effect of the resistance mutations on the binding of the DHPP substrate (and consequently on the binding of pterin site inhibitors) is uncertain, due to the location of the mutations on flexible loops that fall near the *p*ABA binding site, it is unlikely that they would affect the binding of DHPP or any other compounds with affinity for this site. This has direct ramifications on the design of a new class of DHPS inhibitors with affinity for the pterin subsite. Theoretically, agents that inhibit DHPS by binding to this subsite would bypass the resistance mechanisms that have rendered sulfonamide drugs useless for many types of infections. Additionally, the highly conserved nature of the pterin subsite (see section 4.1.1) indicates a possible requirement for many of the pterin subsite residues in catalysis, which may mean that the pterin site would be less likely to undergo resistance conferring mutation.

### 2.1.6 Molecular Dynamics Simulations: Goals and Objectives

In order to gain insight into the catalytic mechanism of DHPS and the conformation of the flexible loops 1 and 2 during catalysis, we performed several series of large scale (2 to 4 nanosecond) molecular dynamics simulations under various solvation conditions using the program AMBER v9 from UCSF.[72,174] The general goals of these simulations were to enable us to visualize the structure of the transition state, deduce the catalytic mechanism of the enzyme, and shed light on the mechanisms of sulfonamide drug resistance. The intent is to use the structures and information obtained from the dynamics simulations to facilitate our design of transition state analogs with strong inhibition of DHPS. A secondary goal of this project was to develop a working model of the DHPS active site, including *p*ABA and DHPP subsites which could be used in subsequent virtual screening experiments.

In this study we performed three series of simulations. The first series of simulations were the substrate/product simulations. This series involved simulations of the apo structure; a ternary complex (Michaelis complex) involving the DHPS enzyme, DHPP, and *p*ABA; and a binary complex involving the pteroate product analog and the DHPS enzyme. The intent of this series of simulations was to allow us to visualize the attack of the nucleophilic nitrogen in *p*ABA on the electrophilic allylic carbon in DHPP and the correct orientation of the attacking *pABA* as it approaches DHPP, with a purpose of further aiding our design of transition state inhibitors.

The second series of simulations were the inhibitor complex simulations which replaced *p*ABA in the first series with sulfamethoxazole in a ternary complex simulation. Additionally, in this series we ran multiple simulations with a sulfamethoxazole-pterin (pterin-SMX) hybrid in the active site with either pyrophosphate or sulfate. The pterin-

SMX compound, Figure 2.8, was synthesized in our lab and crystal structures of the compound in complex with DHPS has been solved but not published. The intent of this series was to gain insight into the unique interactions of sulfonamide drugs with the *p*ABA site and the flexible loops. Of particular interest are the interactions of loop residues with the oxazole ring in sulfamethoxazole, which is very mobile in our crystal structures, and the position of Arg68, whose function during the transition state remains a matter of debate. The goal was to identify these interactions and facilitate the development of tightly binding transition state analogs with potent inhibition of DHPS. Additional insight into the nature of the interaction of the resistance mutation residues was sought during these simulations.

The third and final series of simulations were the resistance mutation simulations. This involved the use of both *p*ABA and sulfamethoxazole ternary complexes with DHPS and DHPP in conditions similar to series 1 and series 3 except that the residues known to cause sulfonamide drug resistance were mutated in these simulations to investigate the mechanism by which they confer sulfonamide resistance. The goal of this final series was to shed light on the mechanisms of resistance and will assist us in our structure-based drug design efforts.

## 2.2 Molecular Dynamics Studies: Materials and Methods

### 2.2.1 Structure Preparation

Several different crystal structures were used as starting points for the simulations described above. For the first series of substrate/product simulations, the pteroate crystal structure, pdb code 1tx0, was utilized. The pteroate product analog present in the crystal structure was replaced with the *p*ABA and DHPP substrates for those simulations, or removed completely for the apo structure simulations. It was left in place for the product simulations. To prepare the structure for the dynamics simulations it was necessary to first generate the initial starting positions for the missing or incorrect residues of loops 1 and 2. This was done by homology modeling the positions of these two loops based upon their known positions in the *M. tuberculosis* and *E. coli* crystal

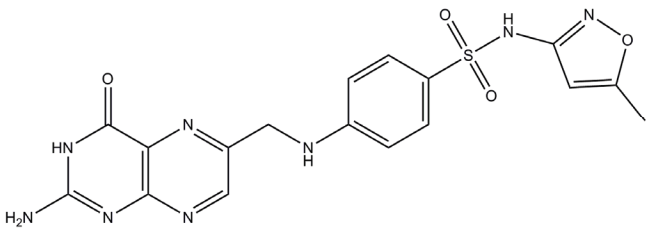



**Figure 2.8. Pterin-SMX Hybrid Compound**

structures. As discussed above, loop 1 is present in the *M. tuberculosis* structure while loop 2 is present in the *E. coli* structure. Both of these loop positions fall near the active site and are theorized to be close to their transition state positions. For loop 1, residues Ile25 through Glu41 were replaced with the corresponding sequence from *M. tuberculosis* and mutated to match the *B. anthracis* sequence. For loop 2, we modeled the positions of residues Gly64 through Val74 on the *E. coli* loop 2 position. This placed the key Arg68 side chain near the anionic binding pocket and allowed it to make an ionic interaction with the sulfate or phosphate substrates. Following this, hydrogens were added to the structure and atoms were typed with Amber atom types. Gasteiger-Huckel charges were used for the substrate and product analogs. A 1000 iteration minimization of the hydrogen atoms was followed by a brief 250 iteration minimization of the flexible loops only to eliminate any steric clash occurring after loop placement. It should be noted that a key active site water was retained in the active site in all dynamics simulations. At this point the structure was ready to take into the simulations phase. Series 3 simulations involving the resistance mutations utilized the same structure, the pteroate analog was replaced with the sulfamethoxazole, *p*ABA, and DHPP structures and the resistance site in question was mutated using the Biopolymer tool of Sybyl.

Simulations involving the sulfamethoxazole-pterin hybrid compound used the pterin-smx crystal structures that had been solved in our previous studies (unpublished data). Two different pterin-smx crystal structures were utilized in these studies. The first structure, like the 1tx0 structure discussed above, was missing much of loop 2, including the position of the key Arg68 residue, and loop 1 was in the incorrect position seen with our previous structures. This structure was prepared in a similar manner as the 1tx0 structure described above. The position of the oxazole ring of the sulfamethoxazole hybrid was not clear in the this crystal structure and it appeared to be able to rotate to an “up” position, tucked between loops 6 and 7, and a “down” in which it is solvent exposed and closer to loops 1 and 2, shown in Figure 2.9, right and left. In the first several simulations that involved the pterin-smx hybrid, the position of the oxazole ring was in the “down” position in the starting structure and the Arg68 side chain occupied the *E. coli* position placing it near the phosphate or sulfate group in the anionic subsite (Figure 2.9, left). During the course of these dynamics simulations a new pterin-smx crystal structure became available that contained more detail regarding the position of the oxazole system as well as more residues in loop 2, including the position of the Arg68 side chain. The remainder of the simulations utilized this starting position (Figure 2.9, right). Only 3 residues from loop 2 were missing from this new pterin-smx crystal structure (Pro69 to Phe71), and they were easily placed and minimized. Importantly, the Arg68 side chain in the newer structure was visualized and appeared to be interacting with the negatively charged sulfonamide group of the pterin-smx ligand. It should be noted, however, that there was no phosphate or sulfate bound in the anionic pocket of this new structure, which may have influenced the position of the Arg68 side chain.

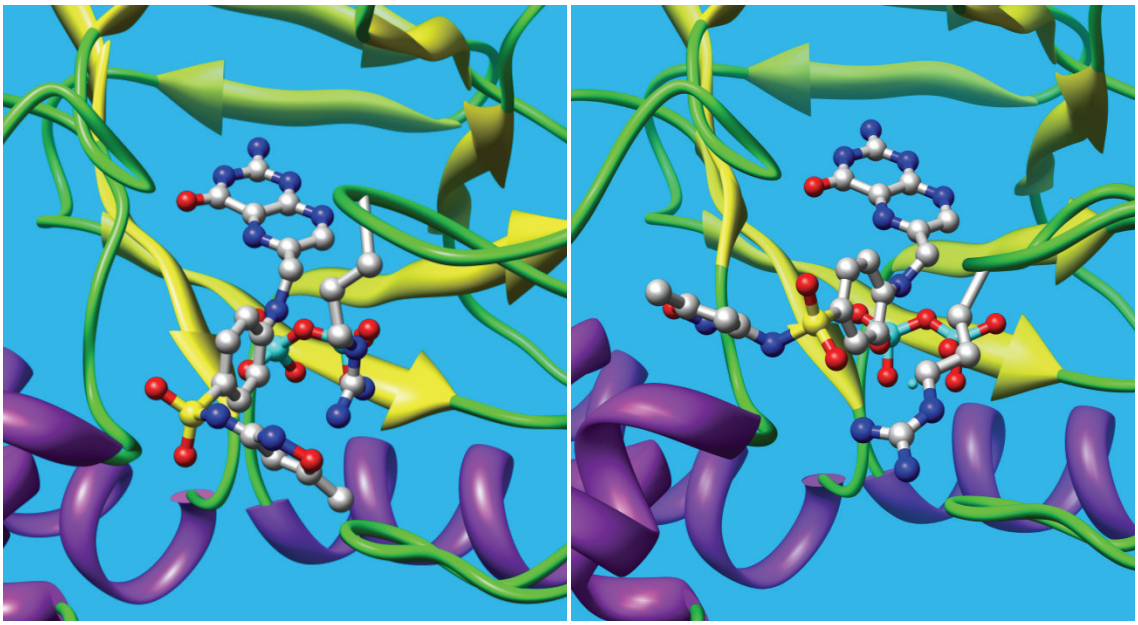


**Figure 2.9. Two Starting Positions for the Pterin-SMX Dynamics Simulations**

### 2.2.2 Force Field and Parameterization

All simulations performed in this study utilized the Amber 2003 force field.[175] Parameters for the standard protein residues in the simulations were generated using the Leap program available in the Amber v9 suite of programs. These parameters have been tested and validated by the Amber developers. It was necessary to develop and load parameters for the non-standard residues in our simulations, including *p*ABA, DHPP, sulfamethoxazole, pterin-smx, pterin-pABA, sulfate, and pyrophosphate. The parameters for the $Na^+$ and $Mg^{2+}$ cations were already available in the Leap program. Parameters for the non-standard residues were generated using the Antechamber program and the General Amber Force Field (GAFF) of Amber v9.[176,177]

Antechamber and the GAFF were specifically designed to develop parameters for organic molecules that are compatible with the traditional Amber force fields and that can be utilized in biomolecular dynamics simulations. Similar to the Amber FF03 (used to load protein parameters), the GAFF uses a simple harmonic function for bonds and angles (see the discussion in Chapter 1 on force field implementation). However, GAFF is much more general than FF03 and covers significantly more organic chemical space. It currently consists of 33 basic and 22 special atom types. The HF/6-31G*, RESP, or AM1-BCC charge methods can be used with the GAFF. In our studies, we used input .mol2 files for each non-standard residue and the AM1-BCC charge method (due to speed and efficiency). Antechamber typed the atoms and bonds, calculated the total number of electrons and net charge, and generated a parameter file for each compound that could be used with the Leap program when developing parameters for the entire

system. Parameters not specifically defined in the GAFF for our non-standard residues were loaded by antechamber based on analogy to similar parameters (after close inspection by the modeler). Appendix A contains the parameter files for each non-standard residue

### 2.2.3 Simulation Methods

A variety of simulation methods were utilized in these studies for each series of simulations. In addition to varying the duration of the simulation, we investigated the effects of implicit versus explicit solvation, presence or absence of cationic cofactor, and presence or absence of the anionic sulfate or phosphate groups. Table 2.3 gives a complete list of every simulation and the simulation design for each series of simulations.

The following general procedure was used to set up a dynamics simulation: First, the program Leap was used to load the parameters for the non-standard residues that had been previously generated as well as all the standard residues in the protein. As mentioned previously, we utilized the Amber 2003 FF for all simulations performed in these studies. The next step was to balance the charge of the system by adding $Na^+$ counter-ions. This was done using Leap for all simulations, both implicit and explicitly solvated. In the case of implicit simulations, the topology and parameter files were then generated. In explicitly solvated simulations, the Leap program was used to generate a 10 Å octahedral solvent box around the system using the TIP3 water model.[127] Topology and parameter files were then saved for these systems.

The Amber module Sander was used to run all minimizations and molecular dynamics simulations. The general procedure for implicitly solvated simulations involved performing a 500 iteration minimization of the system prior to starting the dynamics run (250 iterations using the steepest descent method, followed by 250 iterations using the conjugate gradient method). A non-bonded cutoff of 16 angstroms was used for long range electrostatic interaction calculations in all cases, and the Hawkins, Cramer, Truhlar pairwise generalized Born solvation model was utilized for all implicit simulations.[178] The dynamics simulations that followed were allowed to evolve from 2 to 4 nanoseconds, depending on the simulation. The Langevin thermostat was used to maintain the temperature of the system at 300°K. SHAKE bond length constraints were applied to the bonds involving hydrogen to allow a 2 femtosecond simulation time step.[125] Periodic boundary conditions were not necessary for implicitly solvated systems.

The general procedure for explicitly solvated systems involved an initial 1000 iteration (500 steepest descent, 500 conjugate gradient) minimization of the solvent holding the protein constrained, followed by a 2500 iteration (1000 steepest descent, 1500 conjugate gradient) of the entire system. Non-bonded energy cutoffs of

**Table 2.3. DHPS Molecular Simulations Design Summary**

| Series-Number | PDB[a] | Active Site | Duration | Solvation[b] | Co-factor Cation[c] | Anionic Group[d] |
|---|---|---|---|---|---|---|
| 1-2 | 1tx0 | None (Apo Structure) | 4 ns | Explicit | None | None |
| 1-5 | 1tx0 | pABA, DHPP | 4 ns | Implicit | $Na^+$ | DHPP |
| 1-14 | 1tx0 | pteroate product | 4 ns | Explicit | $Mg^{2+}$ | $SO_4$ |
| 1-16 | 1tx0 | pteroate product | 4 ns | Explicit | None | None |
| 2-1a | n/p | pterin-SMX | 2 ns | Implicit | $Mg^{2+}$ | $SO_4$ |
| 2-1b | n/p | pterin-SMX | 2 ns | Implicit | none | none |
| 2-1c | n/p | pterin-SMX | 2 ns | Explicit | $Mg^{2+}$ | $SO_4$ |
| 2-4 | n/p | pterin-SMX | 3.6 ns | Explicit | none | $SO_4$ |
| 2-6 | 1tx0 | SMX, DHPP | 4 ns | Implicit | $Na^+$ | none |
| 2-13 | n/p | pterin-SMX | 4 ns | Explicit | $Mg^{2+}$ | $SO_4$ |
| 2-15 | n/p | pterin-SMX | 4 ns | Explicit | $Mg^{2+}$ | none |
| 2-17 | n/p | pterin-SMX (oxazole ring in, Arg68 down) | 4 ns | Explicit | $Mg^{2+}$ | PPi |
| 2-18 | n/p | pterin-SMX (oxazole ring out, Arg68 up) | 4 ns | Explicit | $Mg^{2+}$ | PPi |
| 3-3 | n/p | pterin-SMX, F33L | 4 ns | Explicit | none | $SO_4$ |
| 3-7 | 1tx0 | pABA, DHPP, F33L | 4 ns | Implicit | $Na^+$ | DHPP |
| 3-8 | 1tx0 | SMX, DHPP, F33L | 4 ns | Implicit | $Na^+$ | DHPP |
| 3-9 | 1tx0 | pABA, DHPP, T67A | 4 ns | Implicit | $Na^+$ | DHPP |
| 3-10 | 1tx0 | SMX, DHPP, T67A | 4 ns | Implicit | $Na^+$ | DHPP |
| 3-11 | 1tx0 | pABA, DHPP, P69S | 4 ns | Implicit | $Na^+$ | DHPP |
| 3-12 | 1tx0 | SMX, DHPP, P69S | 4 ns | Implicit | $Na^+$ | DHPP |

a. The published pdb code used is indicated when relevant. Non-published, internal crystal structures are indicated by n/p.

b. Explicit solvation used the TIP3 water model in all cases; implicit solvation utilized a Generalized-Born Surface Area Model.

c. The $Na^+$ cation was placed into the anionic site during the charge neutralization step of protein and left in place during simulations. $Mg^{2+}$ atoms, when used, were positioned based upon the 1eye *M. tuberculosis* crystal structure.

d. When DHPP or another pterin-diphosphate analog was simulated in the active site, the diphosphate chain occupied the anionic binding subsite.

10 angstroms were used in all minimization and dynamics steps involving explicitly solvated systems. Periodic boundary conditions were applied during the minimization steps to maintain a constant volume. The Particle Mesh Ewald (PME) summation method was used to calculate long range electrostatic energies in both minimization and dynamics steps for explicitly solvated systems. The minimization steps were followed by a 20 picosecond dynamics simulation which kept weak restraints on the protein while the solvent was allowed to evolve. During this step the temperature was slowly raised from 0 to 300°K using the Langevin thermostat. A constant volume was maintained during this step using the periodic boundary condition. The following step is the full dynamics evolution phase, usually between 2 and 4 nanoseconds in durations. During this step all restraints were removed from the protein and the system was allowed to fully evolve. The temperature was held at 300°K and the pressure was maintained at 1 atomsphere using isotropic position scaling of the periodic boundary. The SHAKE algorithm was employed in all explicitly solvated dynamics simulations to allow for 2 femtosecond time steps.

All dynamics simulations were carried out in parallel using the Linux Cluster at the Hartwell Center of St. Jude Children's Research Hospital. Amber simulations scaled most efficiently to sixteen processors, and this was typically used for a molecular dynamics simulations. Analysis of the completed dynamics simulations was performed using the programs VMD and Chimera.[179,180]

### 2.2.4 Molecular Simulations Analysis

The ptraj analysis tool available in the Amber v9 package was used to perform trajectory analysis in these studies. It should be first noted that full trajectory analysis was only performed against models determined to be stable from visual analysis of the trajectory. Simulations in which the ligands or cofactors were expelled (discussed below) were not considered stable and trajectory analysis was not performed, although observations are made in the discussion regarding key events at specific time points (i.e. ligand expulsion).

Kinetic, potential, and total energy plots were calculated to demonstrate equilibration and stability of the simulation (see Appendix B). Additionally, temperature, pressure, and density plots were used for further demonstration of model stability (data not shown). Dihedral analyses were performed for 4 key binding and mutation residues: Phe33, Thr67, Arg68, and Pro69 (Appendix B). Dihedral analysis was used to determine the degree of conformational sampling for the loop 1 and 2 residues as well as the stability of the residues during the portion of the production phase that was used to determine average structures.

Average and minimum energy structures were calculated and RMSD plots were generated based upon these structures to determine the degree of structural variation and model stability (Appendix B). Average structures discussed below were calculated

from the final 1 nanosecond for simulations extending to 4 nanoseconds and the final 500 picoseconds for simulations extending to 2 nanoseconds. Minimum energy structures were determined from the entire production phase (excluding heating and minimization steps). RMSD values were calculated referenced to the starting structure, the average structure, and the minimum energy structure for the entire protein using backbone atoms, and for both loop 1 and loop2 using backbone atoms and all atoms (Appendix B).

Energy plots, RMSD plots, and dihedral plots determined from dynamics simulation 2-17 (the simulation used to determine the active site model used in our subsequent studies) can be found in Appendix B.

## 2.3 Molecular Dynamics Studies: Results and Discussion

### 2.3.1 Substrate/Product Simulations

The simulations in this series included the DHPS apo structure, a *p*ABA/DHPP/DHPS ternary structure, and a pABA-pterin/DHPS binary structure. The first simulation was a 4ns explicitly solvated simulation of the DHPS apo structure. The starting position for this simulation was essentially the 1tx0 structure with loops 1 and 2 placed as described above, but lacking the pteroate product analog, or any other cofactor. Our anticipation was that during the course of this simulation that the Arg68 residue would insert into the active site and engage in a pi-stacking interaction with the pterin side residue Arg254, as is seen in our apo crystal structure. However, this was not the case. Instead, Arg254 was observed to fold back upon itself and engage in ionic interactions with two Aspartate residues (Asp61 and Asp101) while Arg68 engaged in and maintained ionic interactions with Asp35 on loop1. Figure 2.10 shows the initial configuration of these residues in the starting structure (left) and the interactions that are seen in the final, average structure (right). Although it remains to be determined why the loop positions of the apo crystal structure weren't reproduced in this simulation, one possible explanation is that the simulation did not contain a ligand in the anionic binding subsite, while a sulfate anion was present in this site in the apo crystal structure. As will be discussed below, we have noted in many of our simulations the effect of this negatively charged group on stabilizing the loop positions near the active site. Additionally, we have observed in our crystal structures that Arg254 engages the sulfate (and terminal phosphate of DHPP) in ionic interactions that may stabilize its position in the active site as well.

The second simulation of merit in this series was the *p*ABA, DHPP, DHPS ternary structure. This was a 4 nanosecond implicitly solvated simulation. Of note here is the use of a $Na^+$ cation in the $Mg^{2+}$ site interacting with the diphosphate group of simulation prior to ligand expulsion. The Arg68 residue is engaged in an ionic interaction with the α phosphate of DHPP while Pro69 (a known resistance mutation site)

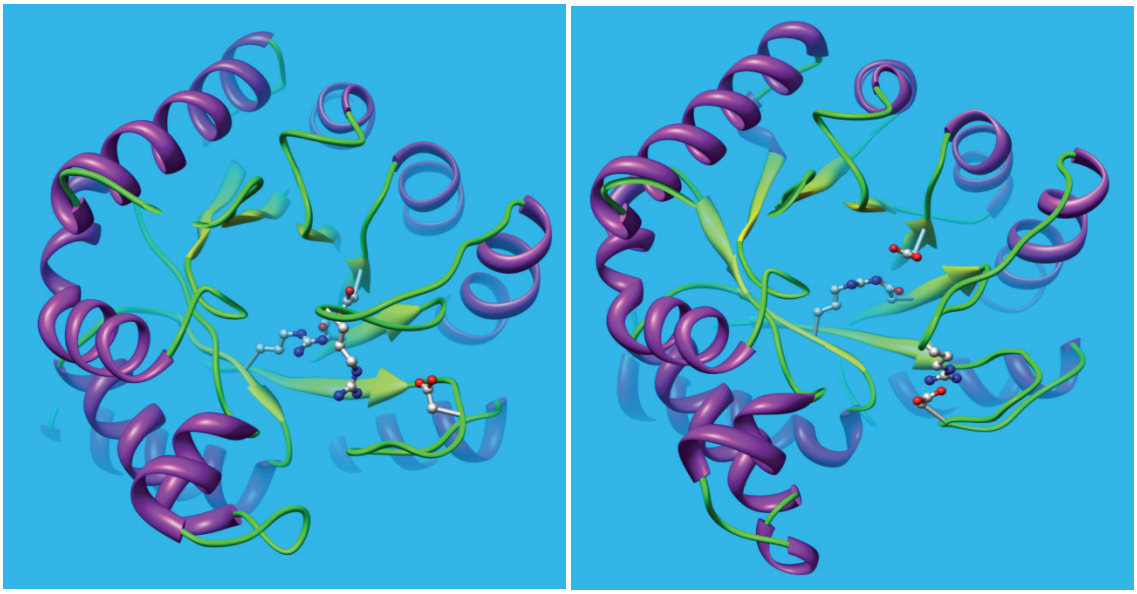


**Figure 2.10. DHPS Apo Simulation Starting and Final Structure**

participates in van der Waals interactions with the *p*ABA substrate. The Pro69 and Lys220 residues facilitate the placement of *p*ABA by forming vdW interactions on both duration of the simulation. In this simulation loop 2 folds completely over the active site and engages in interactions with both the *p*ABA and DHPP substrate. The *p*ABA substrate was ejected from its binding site at 3.2 ns into the simulation, but we were able to note several key interactions prior to this event and gain a clearer understanding of positions of loop 2 during *p*ABA binding. Figure 2.11 depicts the active site of DHPS seen during this simulation with several key interactions highlighted, as determined by calculating an average structure from the 500 ps period of the production phase of the DHPP. This cation was placed by Leap during the charge balancing step of structure preparation and because of its fortuitous position, was left to occupy the site for the sides of the pABA ring. Thr67 (another resistance mutation site) appears to engage in charge-dipole interaction with Lys73 stabilizes possibly helping to stabilize the position of loop 2. Finally, in this simulation we followed Phe33, the third of the conserved resistance sites, and did not observe any direct interaction with the *p*ABA substrate at any time point.

The next two simulations in this series were a set of related, explicitly solvated simulations of the pteroate product analog lasting 4 nanoseconds. The only difference between the two simulations was the lack of a sulfate anion and magnesium cofactor in the second of the two. The purpose of these simulations was to investigate the importance of the anion and cation cofactor in the stabilization of the enzyme-product state and the positions of loops 1 and 2. The differences in these two simulations were dramatic. In the first case (sulfate and cation present) the pteroate product analog

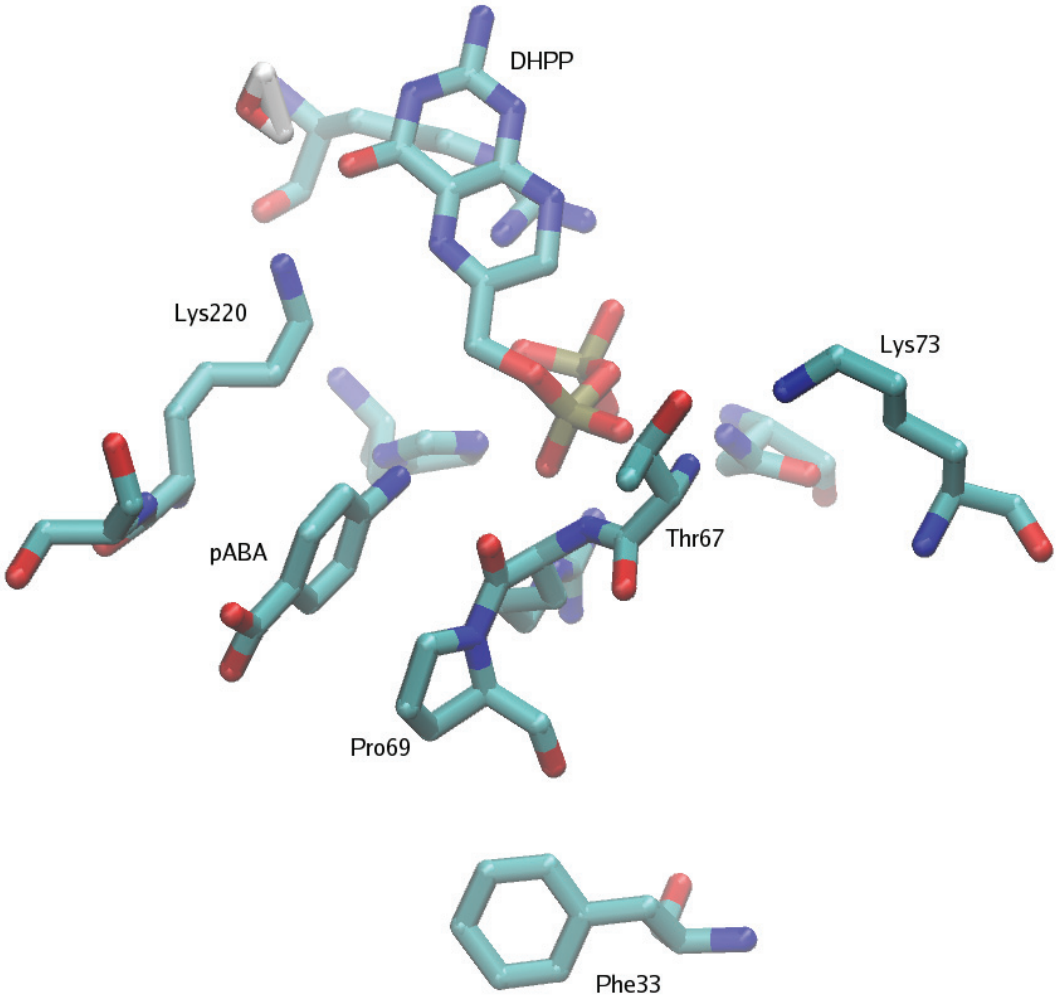



**Figure 2.11. Key *p*ABA and DHPP Active Site Interactions**

maintained its position for the full 4 nanosecond simulation. Arg68 on loop 2 was observed to engage in ionic interactions with the terminal carboxylate of the pteroate compound for the majority of the simulation. In the second case (no sulfate or cation), the pteroate product analog quickly destabilized in the active site and by .6 nanoseconds had begun to be expelled from the active site. Arg254 folded back upon itself in the active site to make ionic interactions with Asp61, similar to what was seen in the apo structure simulations. These results seem to indicate that the position of both loop2 and the active site Arg254 are dependent on the presence of a negatively charged group (sulfate or phosphate), in the anionic binding subsite.

### 2.3.2 Inhibitor Complex Simulations

The next series of simulations involved complexes with the known sulfonamide inhibitor, sulfamethoxazole. The first simulation that we performed was with the ternary sulfamethoxazole, DHPP, DHPS structure, similar to the *p*ABA, DHPP, DHPS ternary complex simulation performed in the first series. Like its corresponding *p*ABA simulation in series 1, this simulation utilized a $Na^+$ ion in the anionic site as the cationic cofactor. Unlike the corresponding *p*ABA simulation, however, in this simulation the sulfamethoxazole was promptly ejected from the active site. This occurred very quickly, by 0.1 ns the sulfamethoxazole was completely removed from the *p*ABA binding site. Interestingly, the DHPP substrate remained in the pterin site for the duration of the simulation. The Arg68 side chain formed and maintained ionic interactions with the terminal phosphate of DHPP for the entire simulation. No interactions with any of the resistance conferring mutations were noted with the sulfamethoxazole ligand in this simulation. The reason for the prompt ejection of sulfamethoxazole from the *p*ABA binding site is unclear as this simulation was set up in exactly the same manner as the corresponding *p*ABA simulation from series 1. Whether this indicates a decreased affinity of sulfamethoxazole for the *p*ABA binding site of the *p*ABA substrate remains to be determined.

Following this we performed a series of simulations involving the pterin-smx hybrid compound that had been developed in our lab (Figure 2.8). The first set of simulations was very similar to the pteroate simulations discussed previously for series 1. These simulations involved a 2 nanosecond implicit simulation of the pterin-smx compound both with and without the $Mg^{2+}$ and sulfate bound in the anionic site. As in the corresponding pteroate simulations, the structure lacking magnesium and sulfate in the anionic pocket was unable to maintain cohesiveness and the pterin-smx compound began to fall out of the active site halfway through the simulation. Again the Arg254 folded back to make ionic interactions with Asp101 and Asp61. The presence of an anionic sulfate or phosphate in this subsite seems crucial to stabilizing the Arg254 residue in its extended form, which appears necessary to keep any pterin substrate in the active site. The structure with the sulfate and magnesium present maintained the pterin-smx in the active site for the duration of the simulation. Arg68 was observed to interact with the negatively charged sulfonamide group for nearly the entire simulation.

The oxazole ring rotated between the “up” and “down” conformations several times during the course of this simulation.

To compare the differences between implicit and explicit solvation and their effects on loop and substrate movements, we designed and ran a pterin-smx simulation, explicitly solvated, for 2 nanoseconds. This simulation included both the magnesium and sulfate groups in the anionic site and corresponded to the 2 ns implicit simulation just discussed. Two key differences were noted with this simulation. First, the oxazole ring did not move between the two positions as quickly as in the implicit solvation simulation. This is likely due to the effect known as “solvent drag”, where the viscosity of the explicit solvent slows small scale movements such as this. The second interesting difference that was noted was that the side chain of Arg68 maintained contact with the sulfate group in the anionic pocket for the duration of the simulation, whereas in the implicitly solvated simulation, the arginine side chain interacted with the sulfonamide group of the pterin-smx hybrid. Figure 2.12 shows the average positions of the pterin-smx, sulfate, cation, and close side chains for the implicit simulation (left) and the explicit simulation (right). The reason for the preference of the arginine for the sulfate in the explicitly solvated simulation is not readily clear, although the altered conformation of Asn27 and the result on the position of the cation is noted and may have played a role in the placement of the arginine side chain.

The final set of simulations that were performed using the pterin-smx hybrid was a set of 4ns, explicitly solvated simulations with pyrophosphate and magnesium in the anionic pocket. We had two goals with this set of simulations. First, to more accurately simulate the “dead-end” product stage of the DHPS reaction by using the pyrophosphate product rather than a sulfate anion in the anionic pocket. Up to this point, difficulties with the parameterization of pyrophosphate had precluded our use of this compound in our

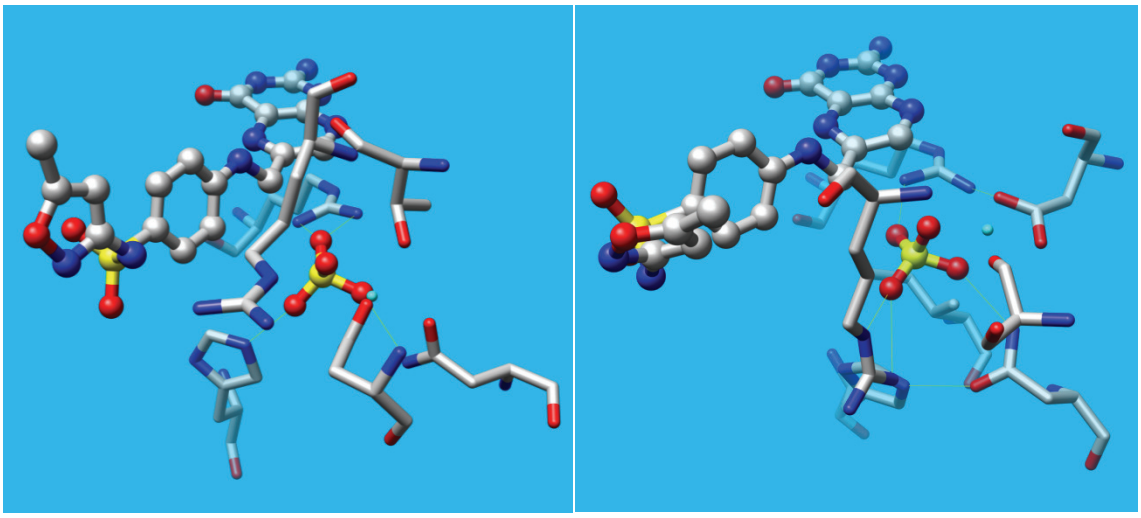


**Figure 2.12. DHPS Pterin-SMX Final Simulation Structures; Implicit Left, Explicit Right**

simulations, but by strengthening the force constants predicted by the Antechamber program, we were able to use pyrophosphate in these simulations. The second goal was to investigate the position of the oxazole ring and the Arg68 side chain. To this end, we utilized the both of the pterin-smx crystal structures that had been solved in our group (discussed above in section 2.2.1). In the first simulation, we utilized the first pterin-smx crystal structure that was solved and placed the oxazole in the "down" position and the arginine side chain interacting with the pyrophosphate. In the second simulation we utilized the second pterin-smx crystal structure as the starting position, with the oxazole ring placed in the "up" position and the arginine side chain interacting with the sulfonamide group. Figure 2.9 shows the starting structure of the active site for both these simulations, oxazole "down" on the left and oxazole "up" on the right. In addition to monitoring the positions of the Arg68 side chain and the oxazole, we also followed the positions of the three conserved resistance conferring mutation sites, Phe33, Pro69, and Thr67, in an attempt to identify any interactions that these residues made with the pterin-smx compound that might be disrupted following a mutation and decrease the binding affinity of the sulfamethoxazole.

In the oxazole "down" simulation the pterin-smx compound remained in the active site for the duration of the simulation, although large movements and loss of some key binding interactions in the pterin subsite were noted midway through the simulation. This corresponded with the loss of the pyrophosphate group from the anionic site at 1.8 ns. The Arg68 side chain maintained an ionic interaction with the pyrophosphate group for the duration of the simulation. In fact, the Arg68 side chain was observed to almost "pull" out the pyrophosphate from the anionic site. Whether this is the normal mechanism for the enzyme's substrate removal or an artifact from the starting position of the simulation remains to be determined. Phe33 and Thr67 were observed to make interactions with the pterin-smx compound and the pyrophosphate, respectively, but Pro69 made no observable interaction with the pterin-smx compound during the simulation period. Although not observed in the average structure calculated, Phe33 was observed to make aromatic stacking interactions with the oxazole ring of the pterin-smx compound at several points during the production phase of the trajectory and seemed to alternate between this position and another stacking interaction with Phe71 in loop 2 (seen in our average structure). The interaction with the oxazole ring may contribute to the stabilization of the sulfamethoxazole compound in the *p*ABA site, while the interaction with Phe71 possibly contributes to the stabilization of loops 1 and 2 in their "active" configurations. The hydroxyl group of Thr67 made hydrogen bonding interactions with the pyrophosphate group and maintained this bond as the pyrophosphate left the anionic site. Figure 2.13 shows the stacking interaction between the oxazole ring and the Phe33 side chain that was observed in this simulation. Also visible is the interaction between the pyrophosphate group and Arg68 and Thr67 as it leaves the anionic subsite (note this is image was not obtained from the average structure).

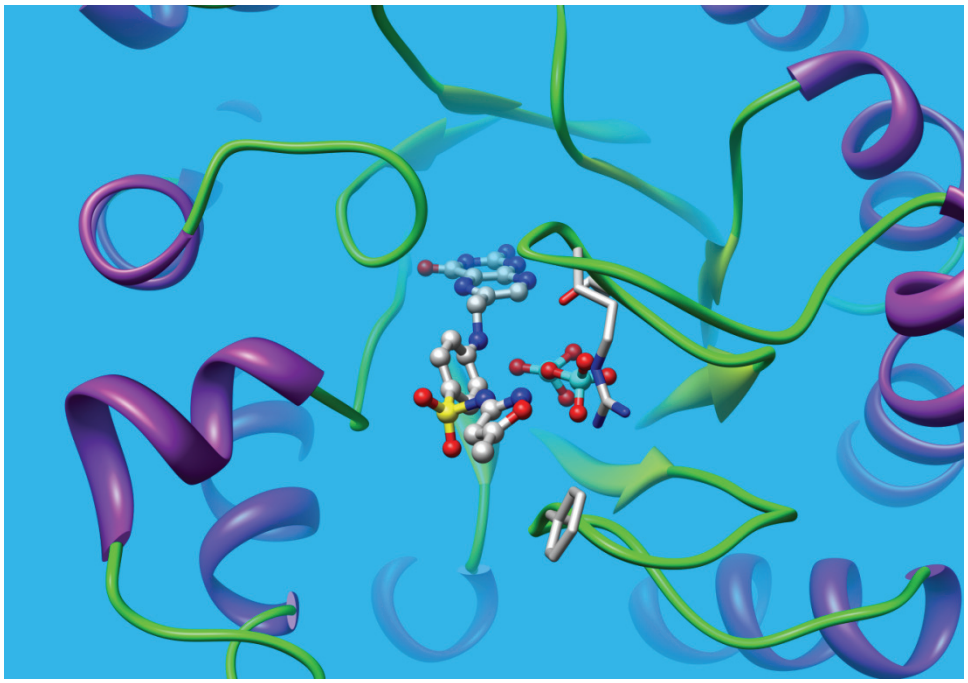


**Figure 2.13. Pterin-SMX Down, 4ns Explicit Simulation**

The matching simulation was performed starting with the oxazole ring in the up position and the arginine side chain interacting with the sulfonamide group of the pterin-smx compound (shown in Figure 2.9). In this simulation the pterin-smx compound stayed tightly bound to the active site and maintained all key binding interactions for the duration of the simulation. The oxazole ring stayed in the up position tucked between loop 6 and loop 7 and was not observed to interact with the Phe33 side chain or any of the other conserved resistance mutation residues. The pyrophosphate group stayed in the anionic subsite for the duration of the simulation. The side chain of Arg68 initially formed an ionic interaction with the negatively charged sulfonamide group of the pterin-smx, but this was quickly lost as the side chain migrated to interact with loop 1 residues, particularly Phe33, shown in Figure 2.14 (average structure).

### 2.3.3 Resistance Mutation Simulations

The final series of simulations that were performed in these studies involved analyzing the effects of the three key mutations, F33L, T67A, and P69S, that have been observed to confer sulfonamide resistance in several bacterial species. In six simulations of 4ns each, under implicit solvation, both *p*ABA and SMX were analyzed in their ternary complex with DHPP with one each of the three mutations applied. In these simulations, a $Na^+$ occupied the anionic subsite along with the diphosphate group of DHPP. These simulations were compared with the corresponding native enzyme

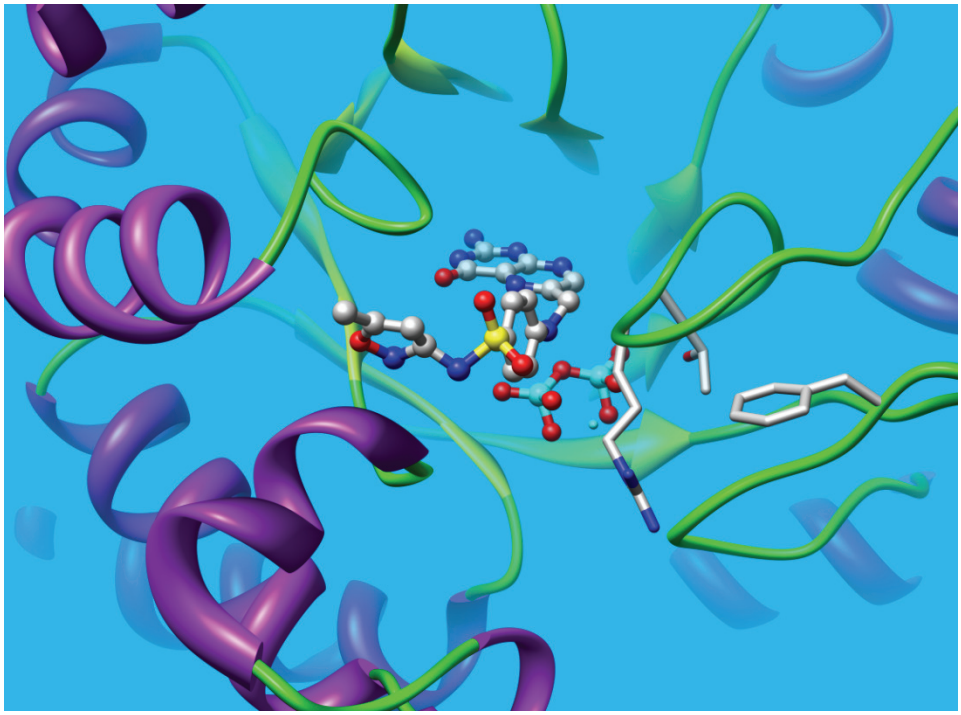


**Figure 2.14. Pterin-SMX Up, 4ns Explicit Simulation**

simulations from series 1 and series 2.

The first set of simulations involved the F33L mutation. In the *p*ABA simulation, the *p*ABA compound was quickly expelled from the active site, by 0.2 ns, no interactions with Leu33 were noted. Arg68 made contact with the terminal phosphate of DHPP, which remained in the pterin site for the duration of the simulation. Loop 1 was observed to lose its position and fold away from the active site. By 2 nanoseconds, loop 1 had begun to form into a helical structure. The reason for the quick loss of the *p*ABA group down position, the sulfamethoxazole compound stayed in the *p*ABA site for the duration of the simulation. Arg68 alternated contact between the negatively charged sulfonamide group and the terminal phosphate of DHPP, at some time points it was able to make a bridging interaction between the two groups. As in the *p*ABA simulation, loop 1 quickly folded out of place and the Leu33 side chain was not observed to interact with sulfamethoxazole at any time. The helix formation of loop 1 was not observed in this simulation. Although the reason for the retention of sulfamethoxazole and the quick expulsion of the *p*ABA compound are not clear, the effects of the F33L mutation on the position of loop 1 over the active site seemed to reinforce our observation from series 2 that the Phe33 residue plays an important role in the stabilization of loop1 during catalysis.

The second set of resistance simulations was the Thr67Ala simulations. In this set of simulations both the *p*ABA and the sulfamethoxazole compounds were

immediately expelled from the active site. No interaction between Ala67 was noted with either *p*ABA, sulfamethoxazole, or DHPP in either simulation. Again, Arg68 made and kept an ionic interaction with the diphosphate group and the DHPP compound remained in the pterin site for the duration of the simulation. Interestingly, in this set of simulations, loop1 was not observed to withdraw from the active site and kept its approximate starting position for the duration of the simulation in both cases.

The final set of simulations was the Pro69Ser mutation. The results of this set of simulations were similar to those seen with the two previous sets. Both *p*ABA and sulfamethoxazole were expelled from the active site early on in the simulations. Arg68 and DHPP behaved in the same manner as that seen in the first two sets of resistance simulations, and loop 1 maintained its position near the active site. Ser69 was observed to make hydrogen bonds to both the *p*ABA and the sulfonamide compounds during and after these compounds were leaving the active site, however the significance of these interactions is not clear as they did not involve the binding of either compound in the *p*ABA subsite.

## 2.4 Summary

The simulations described in this chapter have contributed, at least in part, to our understanding of the binding of both the normal substrates and sulfonamide inhibitors of DHPS. Additionally, we have gained insights into the roles during binding and catalysis of several key residues, whose positions were unclear in our crystal structures as well as the overall positions of loops 1 and 2 during binding and catalysis. In particular the role and position of Arg68, Phe33, Thr67, and Pro69 were closely followed during these simulations. A summary of our findings follows below.

Arg68 has been proposed (and observed in 1 of our crystal structures) to make ionic contact with the negatively charged carboxylate in *p*ABA or the negatively charged sulfonamide group in that class is antibacterials. The majority of our simulations seemed to indicate a preference of the Arg68 side chain for the negatively charged sulfate or phosphate in the anionic binding pocket rather than *p*ABA or sulfonamide group. Although this may be due in part to the starting position of the arginine group, we noted that even when the Arg68 side chain was started in a position where it interacted with the sulfonamide group (as in our “oxazole up” pterin-smx simulations), the contact was not maintained. We noted in one simulation that the Arg68 side chain appeared to facilitate the removal of the pyrophosphate substrate from the anionic pocket, possibly assisted in this by hydrogen bonding interactions made by Thr67.

Phe33 is a key residue that confers resistance to sulfonamide agents when mutated to leucine. We followed this residue closely in all of our simulations and noted an inability of the Phe33 side chain to interact directly with *p*ABA due to distance constraints. However, the aromatic side chain of Phe33 was able to interact with the oxazole side chain of sulfamethoxazole when the simulation was started with this group

out and down. We note the presence of an aromatic group at this position in the majority of the sulfonamide agents that have been marketed. It is possible that this interaction facilitates sulfonamide binding and that the loss of this interaction upon Phe33 mutation decreases the binding affinity of sulfonamide groups for the *p*ABA binding site. This observation can be confirmed by performing activity assessments with a DHPS mutant, sulfamethoxazole (or other aromatic side chain containing sulfonamide), and sulfanilamide (a sulfonamide drug without an aromatic side chain). If this theory is correct, the Phe33 mutation will lower the activity of sulfamethoxazole, but not affect the activity of sulfanilamide. Another possible explanation for the role of Phe33 is the stabilization of loop 1 in a position near the active site during catalysis. We noted in our F33L mutation simulations that loop 1 moved quickly out of position when this residue was mutated, but stayed in position otherwise (as long as a pterin substrate and anion were present in the active site). This phenomenon has been observed by another group, performing similar studies with *S. pneumoniae*.[181]

Our simulations provided us with some insight into the role and function of Thr67 in binding, loop position, and resistance. In the product/substrate simulations Thr67 was observed to make a hydrogen bond with Lys73, also in loop 2, and possibly play a role in the stabilization of loop 2 during catalysis. We also noted, in our pterin-smx simulations, that the Thr67 made hydrogen bond interactions with the pyrophosphate group as it left the anionic pocket, possibly facilitating the removal of this group. We note that the Thr67 side chain is normally not able to interact with the anionic substrate when it occupies its normal position in the anionic pocket. Unfortunately, we did not observe any interactions with Ala67 in the T67A mutation simulations that could provide any insight into the mechanism of this sulfonamide resistance conferring mutation.

The Pro69 residue was observed to play a key role in *p*ABA binding during our product/substrate simulations by making vdW interactions with one face of the pABA ring structure. Presumably, this interaction is disrupted upon mutation to a serine residue such that the sulfonamide agents can no longer bind. However, our mutations simulations with P69S did not reveal the mechanism of this resistance as both compounds were expelled from the active site rather quickly.

In addition to following the residues mentioned above, we were keenly interested in the position of the oxazole ring of sulfamethoxazole during binding and performed simulations with this group in the two positions we observed in our crystal structures. Although, when the oxazole ring was in the “up” position, tucked into a small pocket between loops 6 and 7 it was not able to make the Phe33 stacking interaction, we noted that this position in the pterin-smx simulations was very stable (more so than the “down” position). Another known resistance mutation may lend credence to the oxazole in this position. In N. meningitides, a glycine to cysteine residue has been shown to confer resistance to sulfonamide drugs. This glycine corresponds to Gly188, which resides in the small pocket the oxazole ring fills while in the “up” position. A mutation to cysteine

would presumably block the oxazole ring from occupying this pocket, theoretically leading to a decreased affinity of the sulfonamide drugs for the *p*ABA binding site.

Finally, the role and position of the Arg254 residue and the terminal phosphate of DHPP (or sulfate in several of our studies) deserves mention. This arginine side chain is known (and can be seen in our crystal structures) to play a key role in pterin binding. Our simulations showed in several cases that the position of the Arg254 side chain was dependent on the negatively charged group in the anionic subsite. An extended position was only maintained when the negatively charged group was present. Additionally, even in the presence of a pterin substrate, if the anionic group was absent, the Arg254 side chain withdrew from the pterin site and folded back to make ionic interactions at the back of the pocket. The direct result of this was the destabilization of the pterin substrate in the pterin subsite. This seems to imply that the pyrophosphate product leaves the active site first (perhaps facilitated by Arg68), followed by the pteroate product after Arg254 withdraws from the active site. This observation may have ramification on the design of pterin site binding inhibitors of DHPS.

# CHAPTER 3. MOLECULAR DOCKING VALIDATION STUDIES ON DHPS

## 3.1 Introduction

The overall goal of the research presented in Chapters 2, 3 and 4 is the discovery of novel compounds with significant binding affinities for the pterin pocket of *B. anthracis* DHPS using virtual screening approaches. The pterin binding pocket in DHPS represents an attractive alternative target for the design of novel antibacterial agents. There is a high degree of conservation in the residues that comprise this pocket, and no resistance mutations have been documented in or adjacent to this site, as can be seen from Figure 3.1. To date, a variety of DHPS apo- and holo- crystal structures have been deposited in the Protein Data Bank from six bacterial species (*E. coli*, *S. aureus*, *M. tuberculosis*, *B. anthracis*, *T. thermophilus*, and *S. pneumonia*) as well as one fungal species (*S. cerevisiae*).[151-157] However, prior to embarking on a large virtual screening project against the pterin site of DHPS, it was necessary to investigate different docking and scoring programs and validate their performance. The work presented in this chapter details our extensive docking validation studies against the DHPS pterin site.

### 3.1.1 Why Validate?

Large-scale virtual screening or high-throughput molecular docking (HTD) of in-house or commercial databases has become a common lead discovery technique in drug design. It has been shown to be a complementary tool to traditional, high-throughput screening, with hit rates that can be orders of magnitude higher than those from the latter.[11] In this study, we specifically address the problem of selecting an appropriate docking and scoring combination for virtual screening against a specific target and accurately rank-ordering the virtual hits for further analysis. A review of the literature reveals that there are many docking programs and scoring functions which have been investigated in numerous docking validation studies since 2000.[182-196] It is clear from these studies that, given the large number of docking and scoring functions available, and the variability in their performance with different targets, it is crucial to perform a docking validation study prior to embarking on any virtual screening experiment.[184,185,187,192,193,196] Ideally, the identification of the optimal docking and scoring combination will decrease the number of false positives and false negatives while ensuring optimal hit rates.

### 3.1.2 Docking Validation: Current Methods and Metrics

A number of methods have been reported for validating docking programs and scoring functions.[197,198] One commonly used method is *pose selection* whereby docking

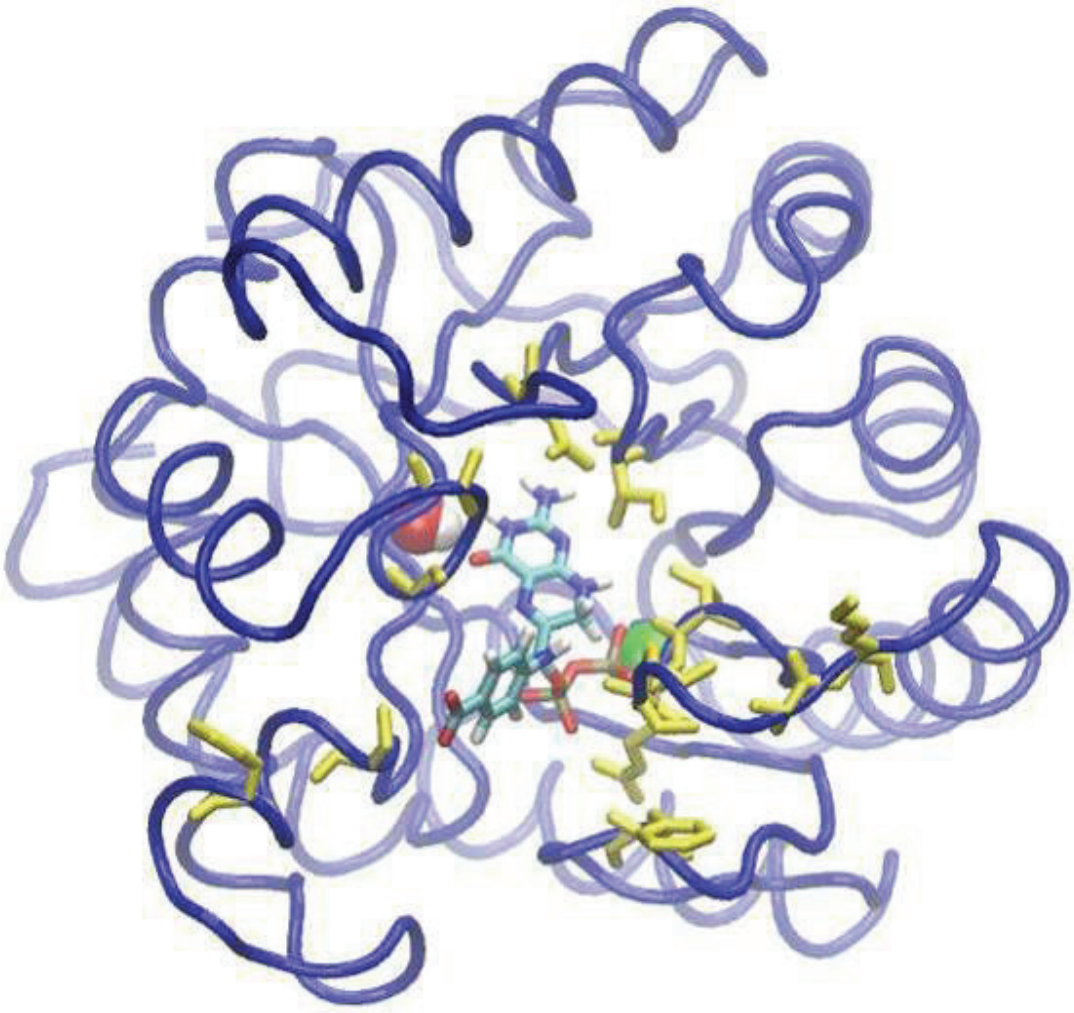


**Figure 3.1. DHPS Structure with Resistance Mutation Sites Highlighted**

Proposed transition state analog shown in active site. Residues conferring sulfonamide resistances are shown in yellow (see Table 2.2).

programs are used to re-dock into the target's active site a compound with a known conformation and orientation, typically from a co-crystal structure. Programs that are able to return poses below a preselected Root Mean Square Deviation (RMSD) value from the known conformation (usually 1.5 or 2 Å depending on ligand size) are considered to have performed successfully. Pose selection is then followed by *scoring and ranking* to study which of the available scoring functions most accurately ranks the poses with respect to their RMSD values.

Another validation method is to dock a so-called *decoy set* of inactive, or presumed inactive, compounds that has been 'seeded' with compounds with known activity against the target in question. After ranking the docked decoy set by score, enrichment can be calculated and enrichment plots or Receiver Operating Characteristic (ROC) curves plotted.[199-201] ROC curves plot the sensitivity (Se) of a given docking/scoring combination against specificity (Sp), and Area's Under the Curve (AUC) can be calculated for comparison. There are two reported advantages of ROC curves over enrichment plots; they are independent of the number of actives in the decoy set and they include information on sensitivity as well as specificity.[198,202] However, the former advantage has recently been challenged.[203]

### 3.1.3 DHPS Validation: Research Project Goals

In this study of the *B. anthracis* DHPS pterin-binding pocket, five docking programs and nine scoring functions were evaluated using pose selection/scoring and enrichment studies. Pose selection and scoring used the 7-amino-3-(1-carboxyethyl)-1-methyl-pyrimido (4,5-c)-pyridazine-4,5(1H; 6H)-dione (**AMPPD**) co-crystal structure, shown in Figure 3.2, as the source structure. AMPPD was first described as a pterin-based DHPS inhibitor by researchers at Burroughs Wellcome Co.[204-207] We have been able to re-synthesize AMPPD and obtain a 2.3 Å resolution co-crystal structure using *B. anthracis* DHPS. RMSD calculations were used to determine how well specific docking/scoring combinations pose and score the ligand in the pterin site. Enrichment studies were performed using 10 compounds also identified in the Burroughs Wellcome efforts, with measured inhibitory activity against *E. coli* DHPS that are known to bind to the pterin-binding site.[206,207] These active compounds were seeded into three separate decoy sets, each of which has been used in previously reported docking validation studies. Enrichment at 1% and 2%, and ROC curves were used to compare docking/scoring combinations, and results across decoy set were also compared.

The work reported here seeks to address eight questions. (1) How useful is simple pose selection and scoring for determining the optimal docking/ scoring combinations for use against a specific target? (2) How do enrichment calculations at 1% and 2% compare with Areas under ROC curves in evaluating the docking/scoring combinations? (3) How important is decoy set selection? (4) How do docking failures affect results and how should these be accounted for? (5) How does post-docking relaxation affect enrichment results? (6) Can the use of consensus scoring improve

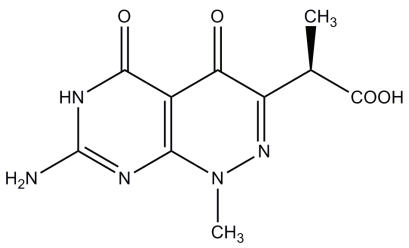



**Figure 3.2. 7-amino-3-(1-carboxyethyl)-1-methyl-pyrimido (4,5-c)-pyridazine-4,5(1H; 6H)-dione, AMPPD**

enrichment results? (7) Is it possible to incorporate the known inhibitory activities of the seeded active compounds to more accurately distinguish between the docking/scoring combinations? Finally, and most importantly for our project, (8) which is the best docking/scoring combination for use in virtual screening against the pterin-binding pocket of the *B. anthracis* DHPS enzyme?

## 3.2 Molecular Docking Validation against DHPS: Methods

### 3.2.1 Docking Programs and Scoring Functions

Five docking programs were evaluated in this study, FlexX, DOCK, Glide, GOLD, and Surflex. FlexX[52] v1.20.1 and Surflex[65] v2.0.1 are included in the Sybyl 7.3 molecular modelling suite of Tripos, Inc.[208] GOLD[57] v3.1.1 was obtained from Cambridge Crystallographic Data Centre (CCDC)[209], Glide[56,210] v4.0 is available from Schrodinger, Inc.[211], and DOCK[49,212,213] v6.0 is freely available to academic institutions from the University of California, San Francisco. FlexX, Surflex, and DOCK use incremental construction algorithms to select compound poses. GOLD uses a genetic algorithm, and Glide is a hybrid method that uses a torsional energy optimization and Monte Carlo sampling[214] for refinement. Nine Scoring functions were investigated. F-Score[52], Surflex-Score[58], ChemScore[63], and GlideScore[56] are empirical scoring functions, PMF-Score[82] is knowledge-based, and D-Score[212], G-Score[57], GOLD-Score and Grid-Score are force-field scoring functions. F-Score, D-Score, G-Score, ChemScore, and PMF-Score are included in the Cscore module of Sybyl 7.3, while Surflex-Score, GOLD-Score, and GlideScore are the native scoring functions for Surflex, GOLD, and Glide, respectively. F-Score is also the native scoring function for FlexX, and Grid scoring was selected for use with the DOCK program.

### 3.2.2 DHPS Target Structure

The crystal structure of AMPPD in complex with *B. anthracis* DHPS, shown in Figure 3.3, was used for all the molecular docking exercises performed in this study. We

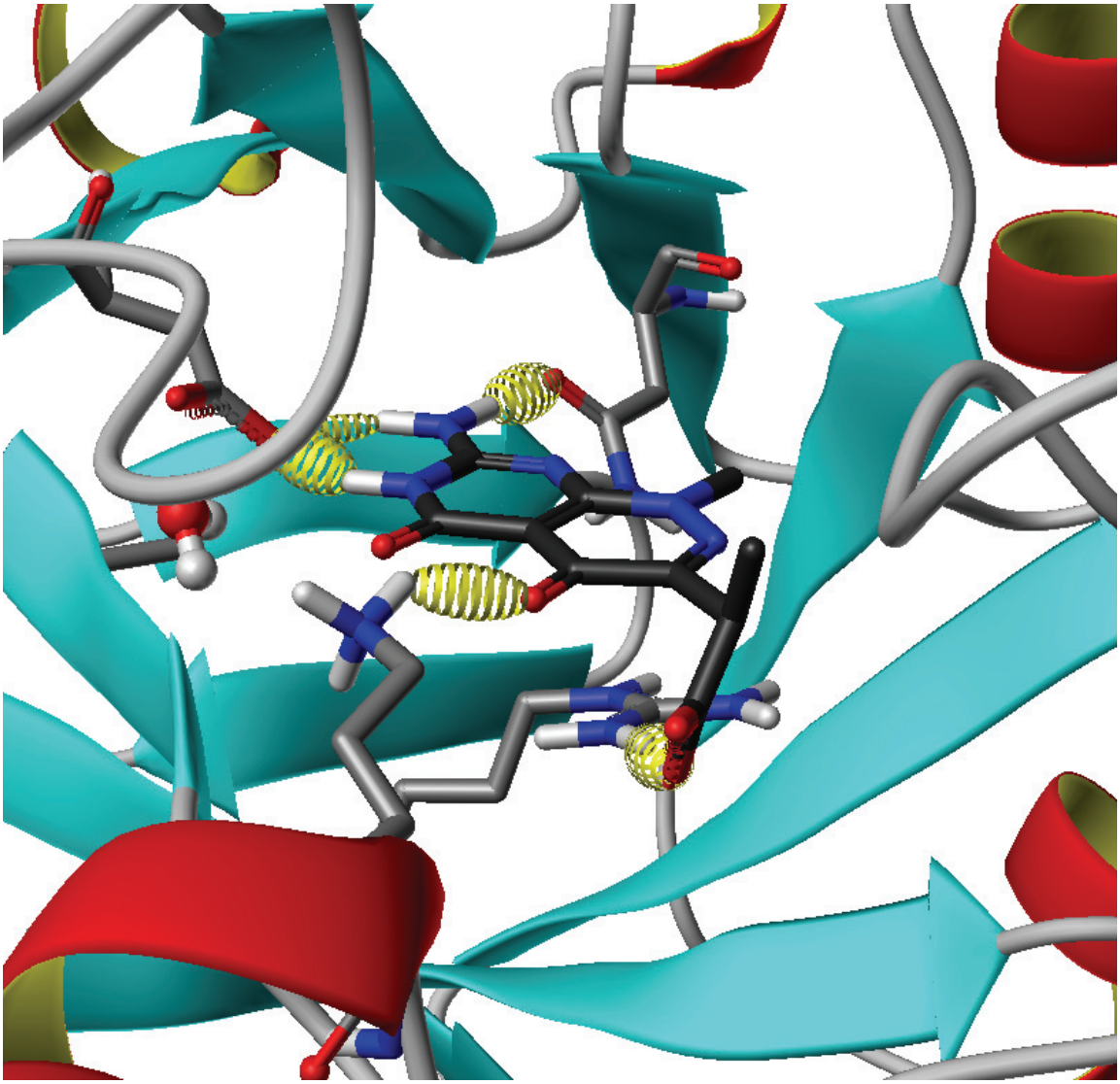


**Figure 3.3. AMPPD Shown Bound into the Pterin Binding Pocket**

Key hydrogen bonds are indicated by spherical ellipsoids.

have determined the structures of *B. anthracis* DHPS in complex with several ligands including pterin site binders and product analogs.[154] The AMPPD structure was chosen for use in this study for three reasons; it binds solely within the pterin binding pocket and does not interact with the adjacent *p*ABA, it has two rotatable bonds which adds an additional degree of complexity to the docking problem compared to the rigid pterin site binders available to us, and its complex with DHPS is one of the highest resolution structures that we have determined. The structure was prepared using the Biopolymer tool of Sybyl 7.3. Missing residues within mobile loops 1 and 2 were modelled using the closely similar *E. coli* and *M. tuberculosis* DHPS structures previously reported and discussed in section.[151,153] Loops 1 and 2 are believed to participate in *p*ABA binding and catalysis, but appear to play little or no role in pterin binding to the enzyme. Hydrogen atoms were added and AMBER FF99 charges were calculated for the protein. A structurally conserved water molecule (WAT1) that interacts with residues Ile187 and Gly216 directly adjacent to the pterin site was included as part of the receptor. A 1000 iteration minimization of the hydrogen atoms was followed by a 100 ps molecular dynamics simulation to refine the positions of the mobile loops 1 and 2. The simulation was performed with the Dynamics tool of Sybyl7.3 using the NTP ensemble, standard temperature and pressure, and 2 fs steps. All residues and ligands with the exception of those in loops 1 and 2 were held under tight constraints. The average structure from the last 20 ps of the simulation was calculated, and a 100 iteration minimization was applied to the entire structure to obtain the final receptor structure.

### 3.2.3 Docking Methodology

***General.*** For consistency, site description files for all docking programs were generated using the AMPPD ligand and an 8 Å spherical radius. WAT1 was included in all docking runs for all programs. ***FlexX v1.20.1.*** A receptor description file was built using the saved .pdb file. Ligands were docked as mol2 files and prepared as discussed below. All other parameters accepted default settings for docking runs. ***GOLD v3.1.1.*** Default speed settings were accepted for both pose selection and enrichment studies. The input structure was the mol2 file with ligand extracted. WAT1 was set 'on' with spin orientation enabled, and the set atom types function was 'on' for ligand and 'off' for the protein. The fitness function was set to GOLD-Score (ChemScore disabled) with default input and annealing parameters. The Genetic Algorithm default settings were accepted as population size 100, selection pressure 1.1, number of operations 100,000, number of islands 5, niche size 2, migrate 10, mutate 95, and crossover 95. All other parameters accepted the default settings. ***Surflex v2.0.1.*** The SFXC file was built using the mol2 prepared protein structure. The protomol was generated using the AMPPD ligand with a threshold of 0.50 and bloat set to 0 (default settings). Ligands were prepared as described below and docked as mol2 files. Cscore calculations were enabled on all Surflex docking runs. All other parameters accepted the default settings. ***Glide v4.0***. The receptor grid was generated using the mol2 file and was based upon the AMPPD ligand and an 8 Å enclosing box. Default values were accepted for van der Waals scaling and input partial charges were used. Standard precision docking was used for

all Glide docking runs, with default settings for all other parameters and no constraints or similarity scoring applied. ***DOCK v6.0.*** The structure and ligand were prepared as discussed above and saved as mol2 files. The molecular surface was generated with the dms tool, included in the DOCK v6.0 package, with a default probe radius of 1.4 Å. Sphgen was used to generate spheres using the dms output and default settings. The active site was defined using the sphere selector tool and an 8 Å radius about the AMPPD ligand, and a corresponding 8 Å grid was generated for scoring using the showbox and grid tools. Flexible ligand docking was utilized with grid scoring as primary and secondary scoring and ligand minimization was enabled. All other docking parameters accepted default settings for docking runs.

### 3.2.4 Ligand Preparation

Ligands were prepared for docking using the Sybyl 7.3 Molecular Modelling Suite of Tripos, Inc. 3D conformations were generated using Concord 4.0[215], hydrogen atoms were added and charges were loaded using the Gasteiger and Marsili charge calculation method.[105] Basic amines were protonated and acidic carboxyl groups were de-protonated prior to charge calculation. The AMPPD ligand was minimized with the Tripos Force Field prior to docking using the Powell method with an initial Simplex[216] optimization and 1000 iterations or gradient termination at 0.01 kcal/(mol*A). Input ligand file format was mol2 for all docking programs investigated.

### 3.2.5 Pose Selection and Scoring

The AMPPD compound was prepared for docking as described above. It was then docked into the DHPS active site of the AMPPD co-crystal structure with each docking program using the methods described above. The number of poses returned by each docking program was determined by the default settings, and the poses were scored using that program's native scoring function. Using the five scoring functions available in the Cscore module of Sybyl, the poses were scored once again in a process that we define as 'rescoring'. The rms analysis tool in the GOLD utilities was used to calculate non-hydrogen RMSD of the docked and scored poses relative to the crystal structure conformation of the AMPPD compound. We used an RMSD of 1.5 Å as our threshold for determining success or failure as opposed to the commonly used 2 Å because of the relatively low number of freely rotatable bonds in the AMPPD compound. For pose selection, the pose with the lowest RMSD was determined from all poses returned by the docking program, regardless of rank. For scoring utility, the RMSD of the best scoring compound was calculated.

### 3.2.6 Enrichment Studies

***Decoy Sets.*** In this study, three compound sets that had been used in previous validation studies were chosen as the decoy sets. The Schrodinger decoy set was used

to validate the Glide docking program.[56,210] It consists of 1000 drug-like compounds with an average molecular weight of 400 D and was downloaded as a 3D SD file from the Schrodinger website. The ZINC decoy set of 1000 compounds was used by Pham & Jain in a validation study of the Surflex scoring function.[217] The Available Chemicals Directory (ACD) decoy set of 861 compounds was used by Bissantz and co-workers in a large docking/scoring validation study.[196] Both the ZINC and ACD decoy sets are available in the Sybyl demo material as 3D SLN files. ***Active Compounds.*** The active compounds that were seeded into each of the decoy sets are shown in Figure 3.4. They were chosen from a previously published series of 65 DHPS inhibitors that are known to bind to the virtually identical pterin site of *E. coli* DHPS.[206,207] The compounds were chosen to reflect as broad a range of binding affinities and structural differences as possible, with the requirement that the activity of the compounds is below an $IC_{50}$ of 20 μM. The compounds were built using the Sketch tool of Sybyl 7.3 and prepared for docking as described above. ***Rescoring***. The highest scoring pose of each compound in the enrichment sets (both active and decoy) was saved for each docking program and imported into a Sybyl Molecular Spreadsheet for rescoring using the Cscore functions F-Score, ChemScore, PMF-Score, D-Score, and G-Score. The effect of relaxing the compounds in the active site using the Cscore relaxation option was investigating by scoring before and after the relaxation. Additionally, a composite score was calculated using the 5 Cscore functions for both the relaxed and unrelaxed scores calculated.

### 3.2.7 Statistical Analysis

We have developed a non parametric statistic, sum of the sum of log rank (SSLR), to test whether a scoring function performs better than random ordering and to compare the docking performances of two scoring functions. The SSLR statistic considers both the ranks of known active compounds relative to the decoy compounds and also the orders of the rank indicated by the $IC_{50}$ values. For a virtual screening experiment, assuming a total of m decoy compounds and n active compounds, the SSLR statistic is defined as:

$$SSLR = \sum_{i=1}^{n} \sum_{j=1}^{i} \log (r_j) \qquad \textbf{Equation 3.1}$$

where $r_j$ is the rank of the jth active compound among all N = m+n compounds; n active compounds are arranged in the order according to their $IC_{50}$ values. By default, the smaller the $IC_{50}$ is the more active is the compound; small SSLR favors early detection of active compounds. ***Test if a scoring function performs better than random scoring.*** The exact distribution of SSLR under null hypothesis is difficult to derive mathematically but can be easily obtained numerically by simulations. The null hypothesis assumes that the ranks of the active compounds are assigned completely at random. We simulate this random scoring study 1 million times and record all their SSLR values. The empirical distribution of the simulated values represents an estimate to the exact distribution. We believe that 1 million simulations should be sufficient enough to produce a reasonably good estimate. The p value of the test is simply the proportion of

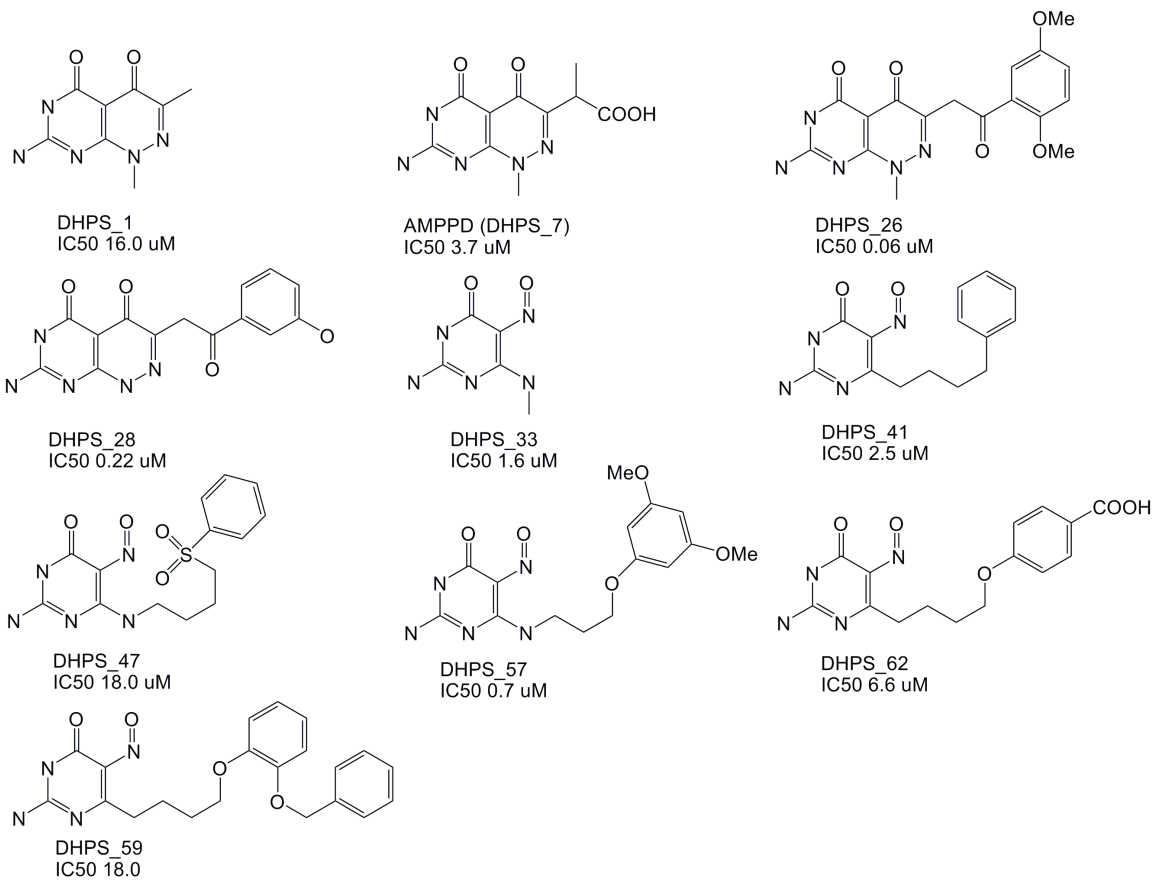



**Figure 3.4. DHPS Active Compounds Used in Enrichment Studies Shown with Activity against *E. coli* DHPS**

the times that the simulated SSLRs are less than the observed SSLR. ***Compare the performances of two scoring functions.*** We have developed a permutation test to compare the performance of two scoring functions. Under the null hypothesis that two scoring function are equal, i.e. $SSLR_x - SSLR_y = 0$, the ranks of the active compounds of the two scoring functions are interchangeable. Assuming $x_i$ and $y_i$ are ranks of the ith active compound for the two scoring functions, the permuted rank is given by:

$$x_i^* = q_i x_i + (1 - q_i) y_i \qquad \textbf{Equation 3.2}$$

and

$$y_i^* = q_i y_i + (1 - q_i) x_i \qquad \textbf{Equation 3.3}$$

where $q_i$ is from Bernoulli distribution with success probability 0.5. Empirical distribution of the difference of *SSLR* is obtained based on the permuted data and the p value of the test is given by the proportion of the times that the permuted differences are greater or less than the observed difference, depending on the direction of alternative hypothesis. ***Missing values.*** In situations where the docking and scoring combination failed to return poses (failed docking), we have penalized the docking/scoring combination by giving those compounds with missing scores the worst score returned by that particular scoring function for a compound in the decoy set (see our discussion on docking failures below).

## 3.3 Molecular Docking Validation against DHPS: Results

### 3.3.1 Pose Selection and Scoring

Table 3.1 shows the results of the pose selection and scoring validation trials. The number of poses returned by the five individual docking programs is listed in parentheses below the docking program name. The best pose, as determined by lowest RMSD, and the rank of that pose by the docking program's native scoring function is given in column 2. Column 3 lists the RMSD of the top scored pose by the native scoring function of each docking program. Scored poses with an RMSD of less than or equal to 1.5 Å are considered to be successful. Each of the five docking programs successfully returned a correct pose, and four of the five native scoring functions ranked a correct pose as the highest. The one exception was the GOLD and Gold-Score function combination which ranked a pose with a 3.29 Å RMSD as the highest. Columns 4 through 8 in Table 1 give the rescoring results with the Cscore scoring functions; the RMSD of the top ranked pose after rescoring is presented together with the rank of that pose by the native scoring function in parentheses. In most cases, the Cscore scoring functions were able to rank successful poses, and the failures are shown in red in Table 3.1. Three of the scoring functions were not able to rank the FlexX poses, and D-Score

**Table 3.1. Pose Selection and Scoring Results**

| Docking Program | Best Pose (Pose Rank) | Native Scoring Function (1st) | F-Score (Pose Rank) | G-Score (Pose Rank) | D-Score (Pose Rank) | ChemScore (Pose Rank) | PMF-Score (Pose Rank) |
|---|---|---|---|---|---|---|---|
| **FlexX** (30 Poses) | 0.56 Å (3rd) | 1.19 Å | 1.19 Å (1st) | 1.87 Å (10th) | 17.22 Å (21st) | 1.87 Å (10th) | 1.03 Å (2nd) |
| **Surflex** (10 Poses) | 1.48 Å (3rd) | 1.49 Å | 1.50 Å (10th) | 1.50 Å (10th) | 1.49 Å (6th) | 1.49 Å (1st) | 1.48 Å (4th) |
| **Glide** (30 Poses) | 0.37 Å (10th) | 1.13 Å | 0.40 Å (11th) | 0.43 Å (15th) | 1.13 Å (1st) | 0.37 Å (10th) | 0.40 Å (12th) |
| **GOLD** (5 Poses) | 1.30 Å (4th) | 3.29 Å | 1.30 Å (4th) | 3.37 Å (3rd) | 1.30 Å (4th) | 1.30 Å (4th) | 1.30 Å (4th) |
| **DOCK** (10 Poses) | 0.38 Å (10th) | 0.85 Å | 0.85 Å (1st) | 1.00 Å (4th) | 0.52 Å (10th) | 0.85 Å (1st) | 1.00 Å (5th) |

Non-hydrogen RMSD values are shown; RMSD values less than 1.5 Angstroms are considered correct poses, greater than 1.5 Angstroms are considered failed.

performed particularly poorly in this respect, ranking a failed docking pose (outside the pterin site) as the highest. However, D Score was able to correctly rank a successful G-Score and ChemScore were not able to correctly rank poses generated by FlexX, and G-Score also failed with the poses generated by GOLD. Overall, F-Score and PMF-Score correctly rescored the poses generated by all the docking programs and were the best performing functions in this respect. We also note that the poses returned by the Surflex, Glide, and DOCK programs were always successfully scored by both the native functions and the Cscore functions.

### 3.3.2 Enrichment Studies

Figures 3.5, 3.6, and 3.7 show the calculated enrichment at 1% for each docking program/scoring function combination when used with each of the three decoy sets used in this study. It should be noted that we were not able to complete the GOLD docking of the ACD decoy set due to licensing issues, but the ZINC and Schrödinger decoy sets were successfully docked by the GOLD docking program. Enrichment is defined as the number of active compounds detected at a given percent of total decoy set by score ranked pose. Enrichment was calculated at 1% and 2% of the total decoy set rather than 1% and 2% of compounds successfully docked. This requires further explanation. Table 3.2 displays the number of poses (1 pose per compound) returned by the docking programs investigated in this validation study. It is apparent that some programs were able to return more poses than other programs, and this must be taken into account so as not to unfairly penalize programs that failed to dock some of the decoy compounds.

Several observations can be made from the data presented in Figures 3.5 to 3.7. First, the two force field based functions, D-Score and G-Score, and the empirical function ChemScore all performed poorly for each decoy set. Second, the Glide and Surflex docking programs with their native scoring functions performed well (4 or more actives detected at 1%) against each of the three decoy sets. Finally, when used as the FlexX native scoring function, F-Score performed poorly against all three decoy sets, but when used to rescore for the other four docking programs F-Score returned modest to good results. Most notably, F-Score detected 5 of the 10 active compounds when used with DOCK against the ZINC validation set.

Enrichment was also calculated at 2% of the total decoy set docked for comparison (see supplementary material). D-Score, G-Score and ChemScore continued to perform poorly. The scoring functions F-Score and PMF-Score were able to detect on average 1 or 2 more active compounds at 2%. Notably, the top performers at 1%, GlideScore and SurflexScore, continued to show excellent results at 2%, detecting between 6 and 8 of the 10 active compounds.

When comparing the enrichment results with respect to the choice of decoy set, there was a clear difference in performance for the various docking/scoring combinations. Overall, the ZINC decoy set returned the best enrichment results, while

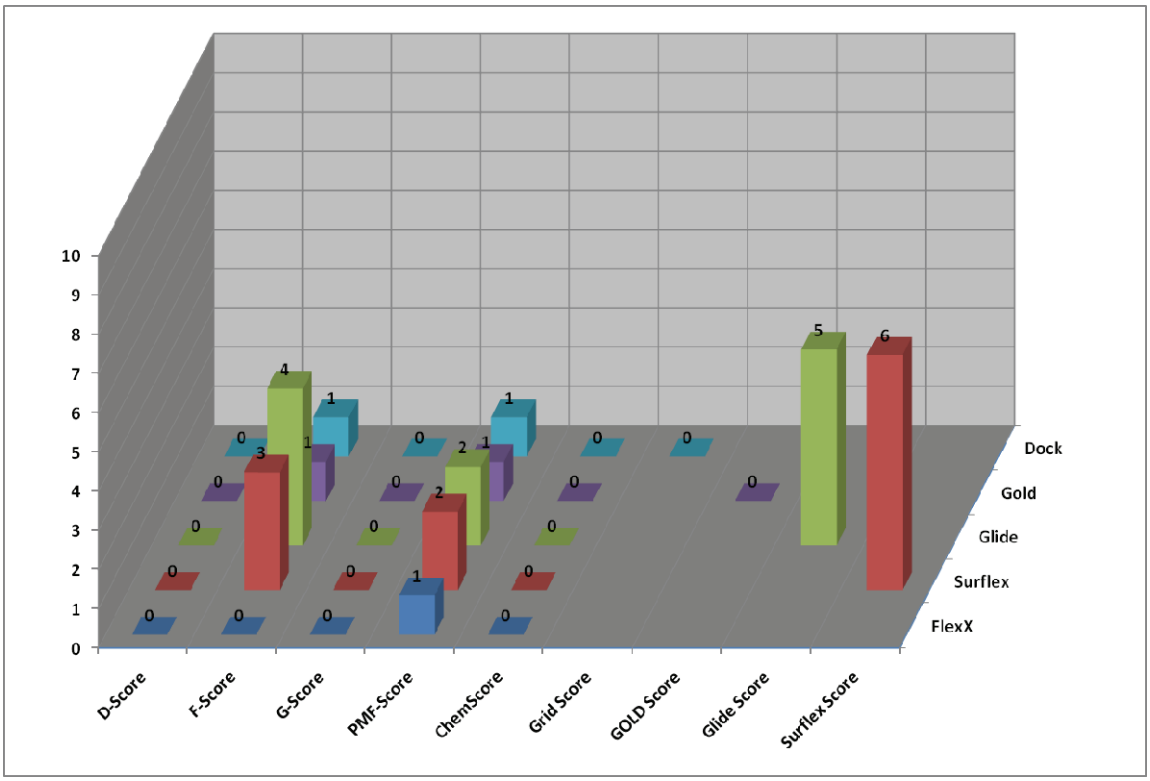



**Figure 3.5. Enrichment Factors at 1% of Total Validation Set Docked, ZINC Decoy Set**

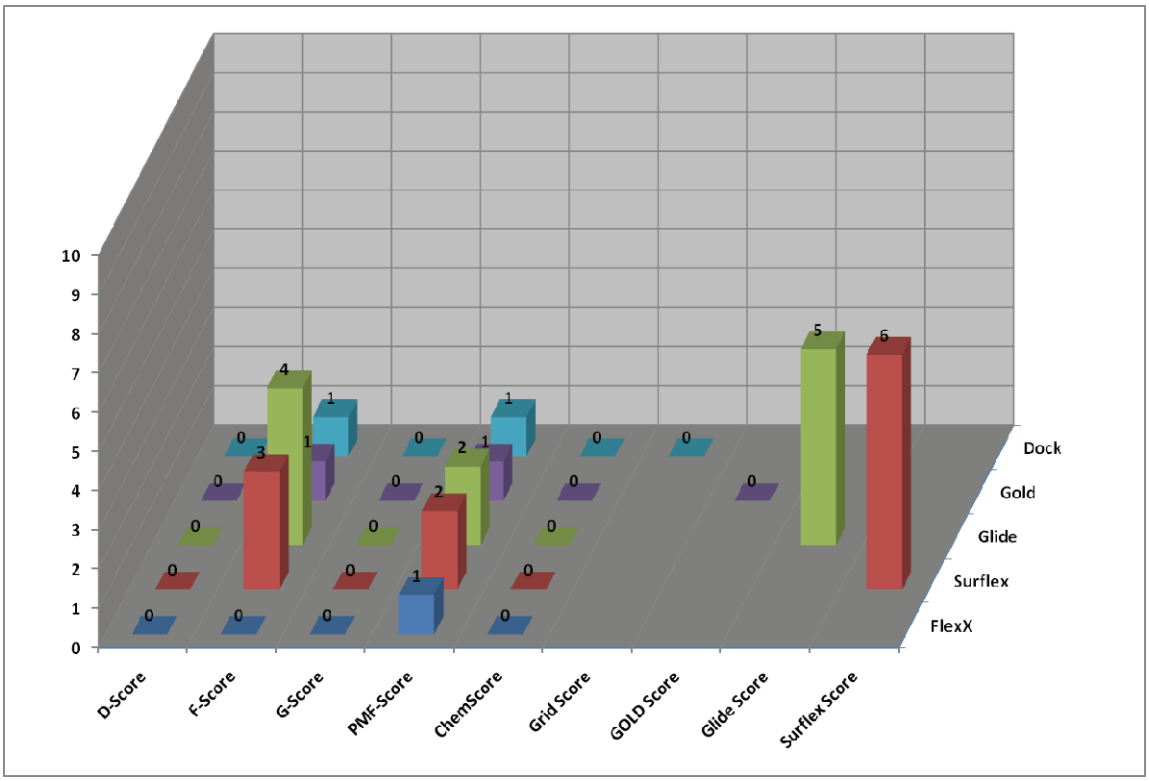



**Figure 3.6. Enrichment Factors at 1% of Total Validation Set Docked, Schrödinger Decoy Set**

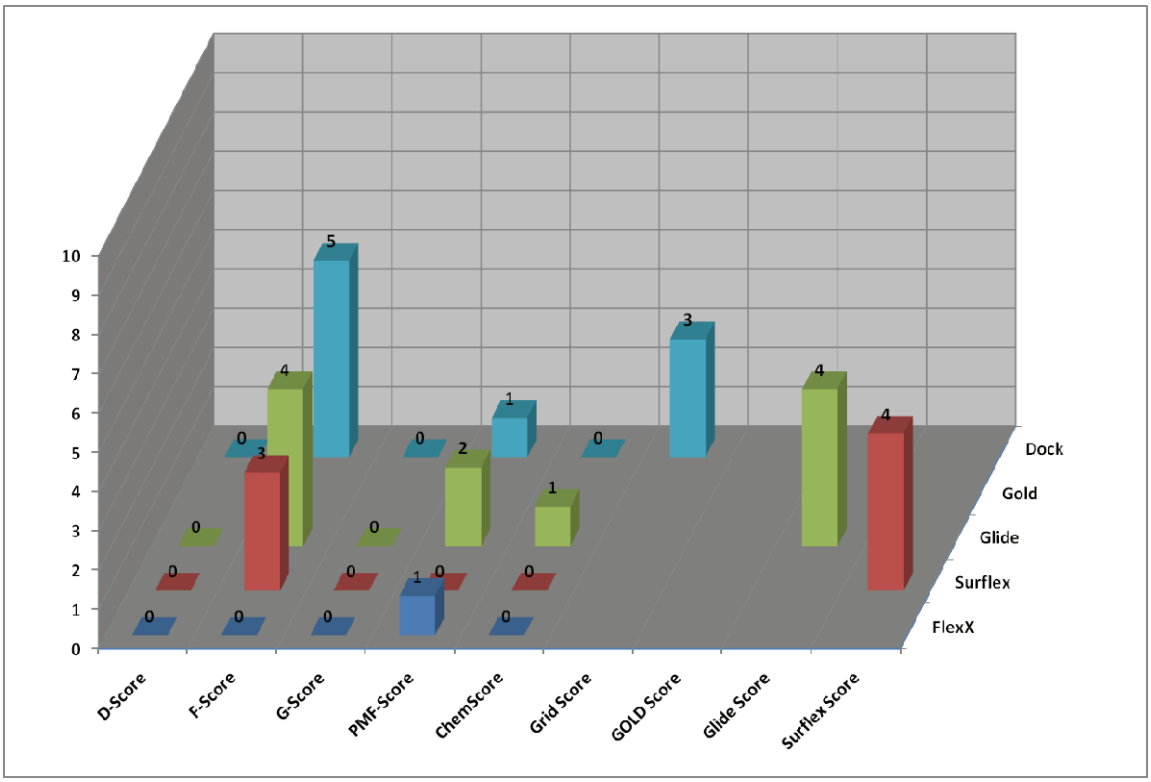



**Figure 3.7. Enrichment Factors at 1% of Total Validation Set Docked, ACD Decoy Set**

**Table 3.2. Number of Compounds Docked by Validation Set**

| Docking Tool | ACD | Schrödinger | ZINC |
|---|---|---|---|
| Full Set | 871 | 1010 | 1010 |
| DOCK | 749 | 752 | 852 |
| FlexX | 811 | 991 | 1004 |
| Surflex | 870 | 1010 | 1010 |
| GOLD | n/d | 1010 | 991 |
| Glide | 579 | 607 | 819 |

the ACD decoy set returned the worst results. It might be expected that the docking programs would have the most difficulty in distinguishing the active compounds from the decoy set when they are close in size and lipophilicity, but this trend was not seen in our enrichment studies. The Schrödinger decoy set differed most from the active compounds with respect to these two parameters but returned the worst enrichment results, while the ZINC and ACD sets, which have the closest parameters, yielded better enrichment results.

### 3.3.3 Receiver-Operating Characteristic Curves

Figures 3.8, 3.9 and 3.10 show representative ROC plots for three of the five docking programs evaluated in this study (see Appendix C for additional ROC plots). The results from the native scoring functions and from rescoring with the five Cscore scoring functions are shown. The calculated areas under the receiver-operating characteristic curves (AU-ROC) values for each docking program with its native scoring function and the five Cscore functions are given in Table 3.3, and are color coded according to performance; green - excellent (above 0.9), black – moderately well (0.9 to 0.6), red – poor (less than 0.6). Calculated *p* values are shown in parentheses in Table 3.3. At a significance level (α) of 0.05, *p* values less than 0.05 indicate significant improvement over random selection while *p* values greater than 0.05 indicate no significant difference over random selection. It should be noted that, when creating the ROC curves, we used the total number of compounds in the validation set rather than total number of docked compounds to enable a more direct comparison of the performance of the docking and scoring algorithms. This point has been discussed earlier with respect to enrichment values, but it is also relevant here. As can be seen from Figure 3.11, when calculating the area under the ROC for Glide using both the total Schrödinger decoy set versus the total successfully docked, there is a small but noticeable difference. This presents a problem when comparing results with a docking program such as Surflex that was able to dock the complete Schrödinger decoy set.

Four observations can be made from these results. First, unlike the enrichment results at 1% and 2%, there is little difference in the ROC results for docking programs when compared across decoy sets. Generally, when a docking/scoring combination

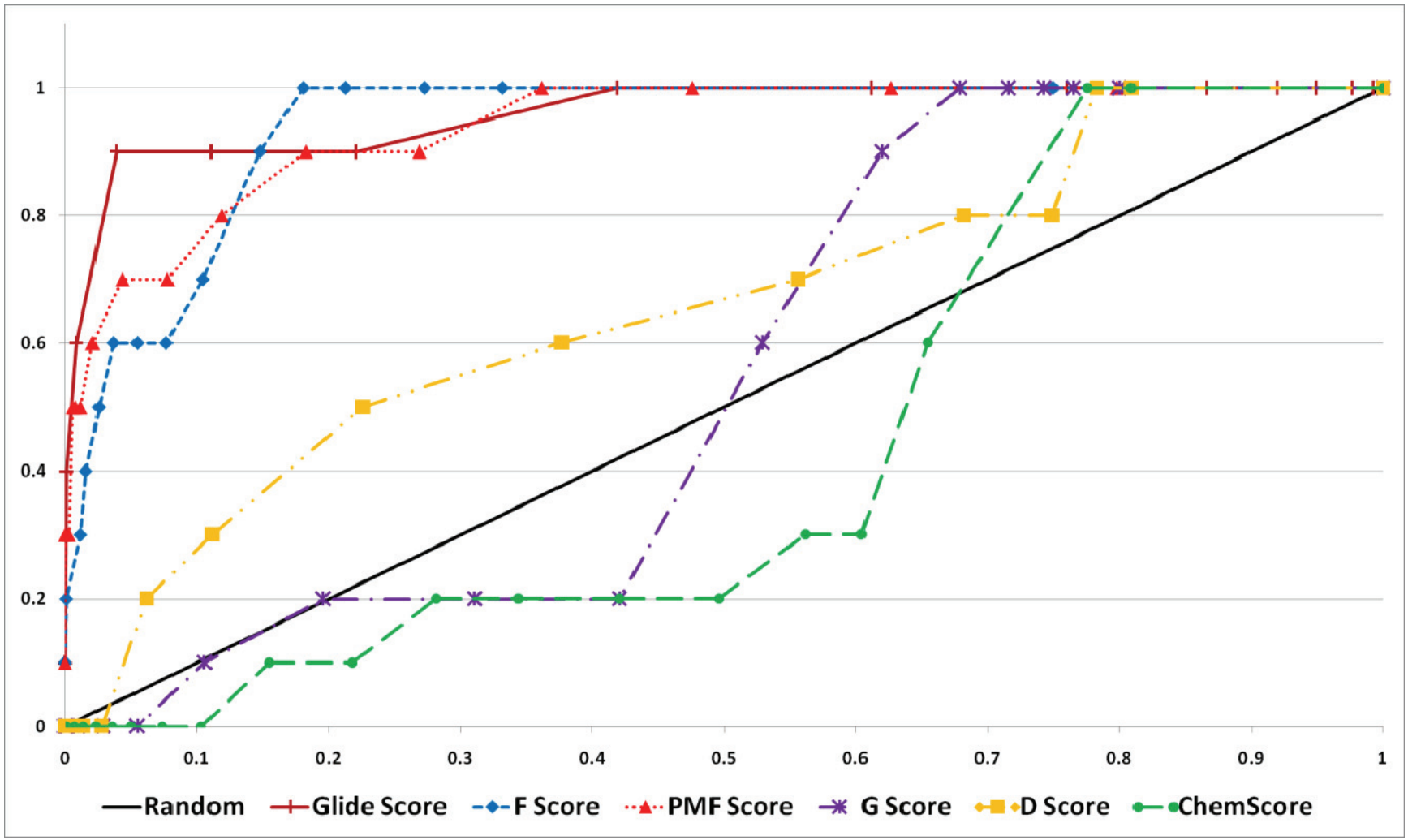



**Figure 3.8. Selected ROC Plots: Glide Docking of ZINC Decoy Set**

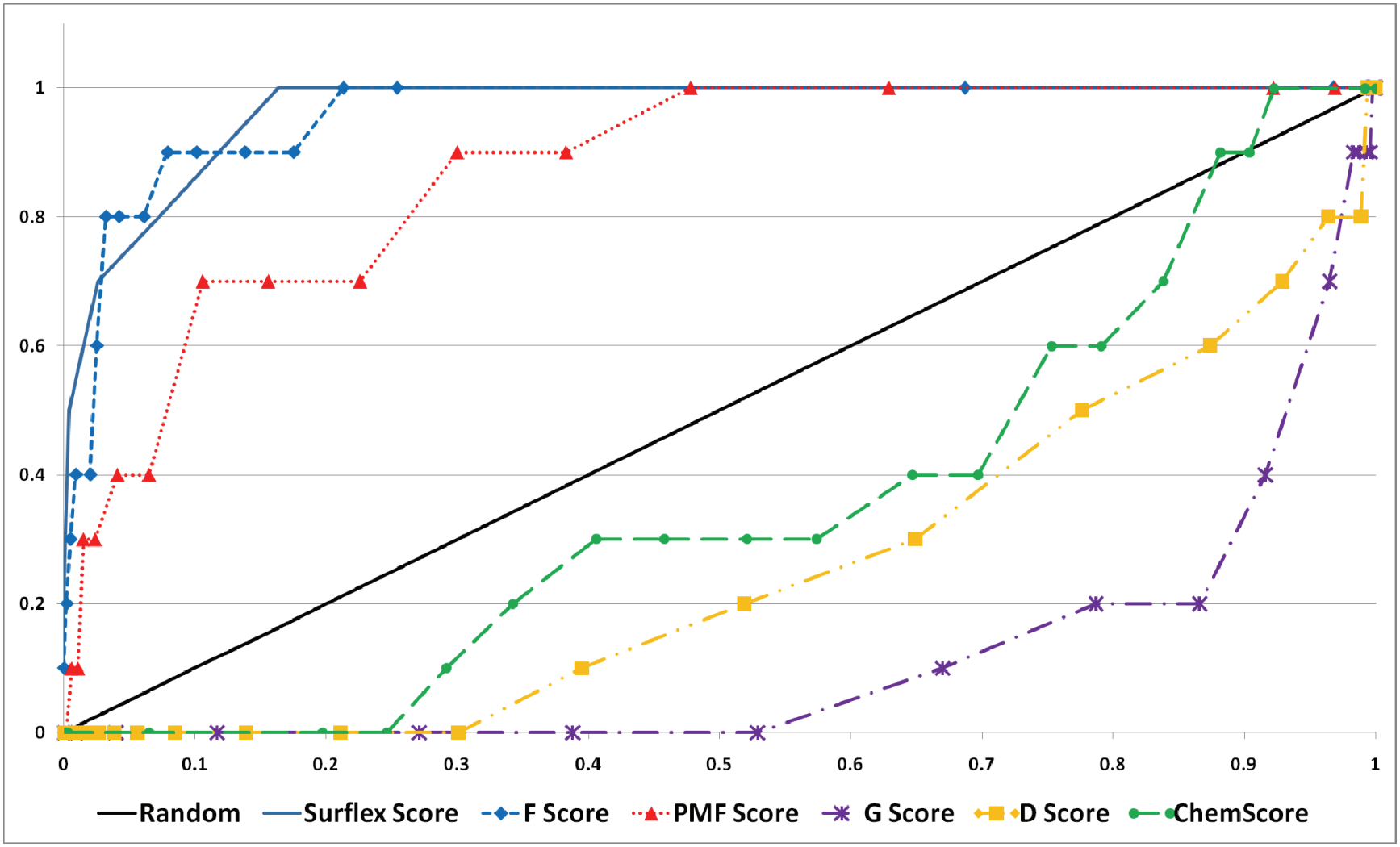



**Figure 3.9. Selected ROC Plots: Surflex Docking of Schrodinger Decoy Set**

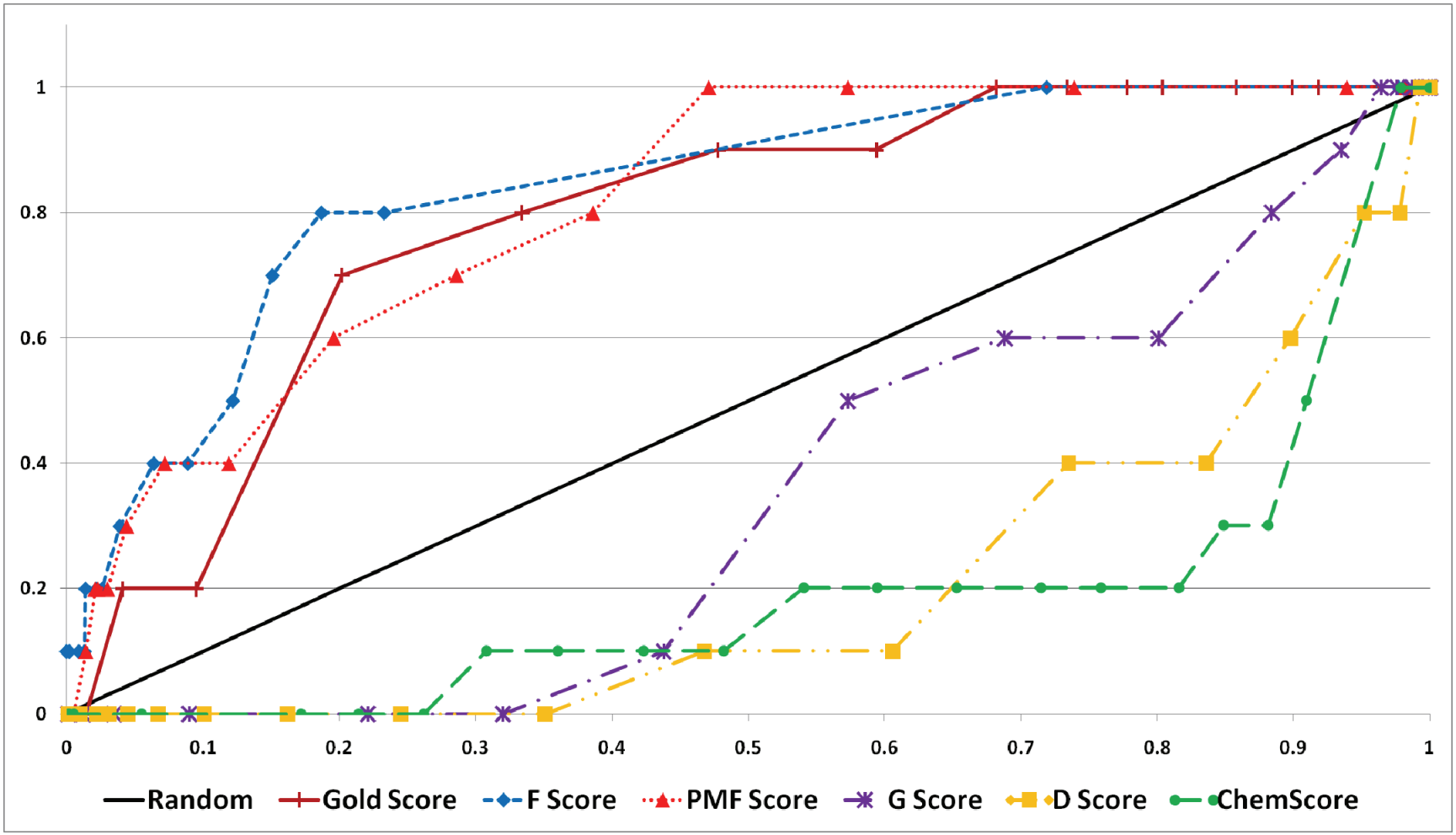



**Figure 3.10. Selected ROC Plots: GOLD Docking of Schrodinger Decoy Set**

**Table 3.3. Calculated AU-ROC with p Values from ROC Curves for 5 Docking Programs, Native Score and Cscore Functions (Unrelaxed)**

| Docking Program / Validation Set | Native Score[a] | F-Score | PMF-Score | G-Score | D-Score | ChemScore |
|---|---|---|---|---|---|---|
| DOCK - ZINC | 0.835 (<.001) | 0.804 (.001) | 0.962 (<.001) | 0.533 (.338) | 0.721 (.007) | 0.540 (.297) |
| DOCK - Schrödinger | 0.770 (<.001) | 0.793 (.001) | 0.932 (<.001) | 0.584 (.110) | 0.689 (.008) | 0.538 (.271) |
| DOCK - ACD | 0.902 (<.001) | 0.860 (<.001) | 0.958 (<.001) | 0.652 (.021) | 0.794 (<.001) | 0.633 (.010) |
| FlexX - ZINC | F-Score | 0.813 (<.001) | 0.932 (<.001) | 0.394 (.854) | 0.588 (.183) | 0.317 (.998) |
| FlexX - Schrödinger | F-Score | 0.746 (<.001) | 0.889 (<.001) | 0.386 (.887) | 0.528 (.376) | 0.289 ( .999) |
| FlexX - ACD | F-Score | 0.891 (<.001) | 0.915 (<.001) | 0.491 (.534) | 0.701 (.006) | 0.506 (.461) |
| Glide - ZINC | 0.971 (<.001) | 0.941 (<.001) | 0.939 (<.001) | 0.558 (.182) | 0.666 (.039) | 0.547 (.267) |
| Glide - Schrödinger | 0.982 (<.001) | 0.947 (<.001) | 0.889 (<.001) | 0.709 (<.001) | 0.654 (.004) | 0.651 (.010) |
| Glide - ACD | 0.977 (<001) | 0.975 (<.001) | 0.936 (<.001) | 0.588 (.029) | 0.738 (<.001) | 0.728 (<.001) |
| Surflex - ZINC | 0.985 (<.001) | 0.980 (<.001) | 0.956 (<.001) | 0.189 (>.999) | 0.436 (.755) | 0.506 (.472) |
| Surflex - Schrödinger | 0.978 (<.001) | 0.963 (<.001) | 0.880 (<.001) | 0.117 (>.999) | 0.251 (.999) | 0.360 (.966) |
| Surflex - ACD | 0.975 (<.001) | 0.975 (<.001) | 0.926 (<.001) | 0.221 (>.999) | 0.467 (.661) | 0.508 (.448) |
| GOLD - ZINC | 0.763 (.002) | 0.923 (<.001) | 0.930 (<.001) | 0.398 (.883) | 0.401 (.862) | 0.237 (>.999) |
| GOLD - Schrödinger | 0.778 (<.001) | 0.846 (<.001) | 0.827 (<.001) | 0.345 (.993) | 0.197 (>.999) | 0.185 (>.999) |

*a.* *p* values <.05 for AU-ROC values >.5 indicate statistically significant improvement over random selection, *p* values <.05 for AU-ROC values <.5 indicate statistically significant decrement over random selection.

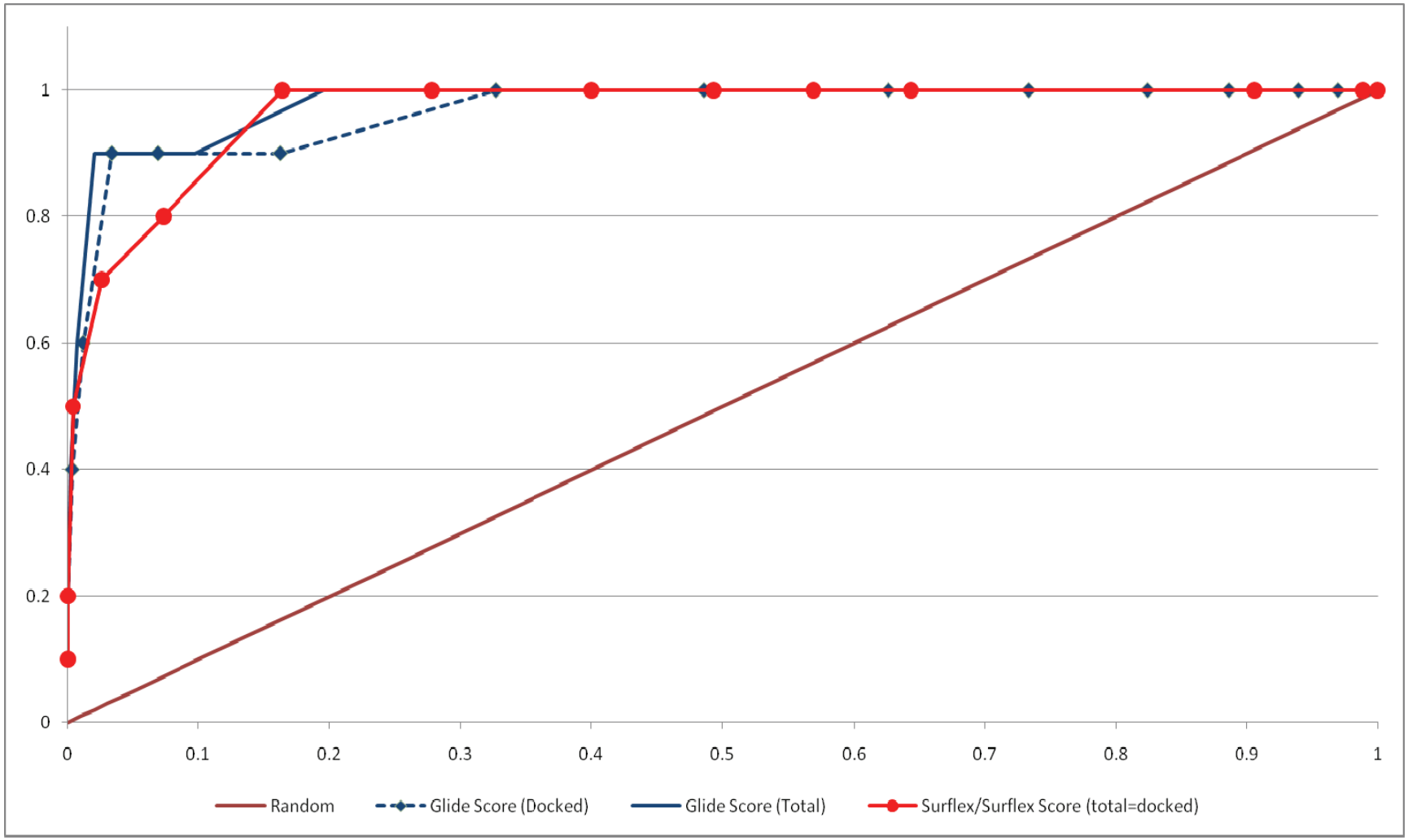



**Figure 3.11. ROC Comparison of Docked versus Total Set, Schrodinger Decoy Set**

performed well, it did so against all three decoy sets. The opposite is also true, with poorly performing docking/scoring combinations consistent with all three decoy sets. One exception to this is the noticeable (although not statistically significant) decrease in performance of PMF-Score when rescoring docking output for the Schrödinger decoy set. We also noted the improvement in performance of the D-Score function when rescoring DOCK output and attribute this to the fact that the D-Score function is based on the original DOCK scoring function by Kuntz, et al.[212] Second, docking programs generally performed moderately to very well when paired with their own native scoring functions. Glide with Glide-Score and Surflex with Surflex-Score performed exceptionally well, and no improvement to the AU-ROC values was seen when rescoring these poses. DOCK, FlexX, and GOLD performed moderately well when scored with their native scoring functions, Grid-Score, F-Score, and GOLD-Score, respectively, but these showed significant improvement upon rescoring. Specifically, AU-ROC values were markedly improved when DOCK results were rescored with PMF-Score, FlexX results with PMF-Score, and GOLD results with both F-Score and PMF-Score. Third, F-Score and PMF-Score generally performed well in rescoring. Curiously, F-score only performed moderately well with its partner FlexX, but performed exceptionally well when used to re-score the outputs of Glide, Surflex, and GOLD. Finally, we note the moderate to poor performance of G-Score, D-Score, and ChemScore when these functions were used to re-score docking output from all five docking programs. Their performance ranged from moderate with DOCK and Glide, to exceptionally poor with Surflex and GOLD.

### 3.3.4 SSLR Calculations

The SSLR value reflects the ability of the docking and scoring combination to detect active compounds early and also their ability to correctly rank the active compounds according to their known inhibition constants. Table 3.4 shows the calculated SSLR statistic and *p* values for each of the docking/scoring combinations evaluated in this study. Lower values for SSLR are more desirable, and *p* values (shown in parentheses) of less than 0.05 indicate that the particular combination showed significant improvement over random selection and ordering. Like the AU-ROC values, the SSLR values demonstrate a clear distinction between the performance of the native scoring functions, F-Score, and PMF-Score over G-Score, D-Score, and ChemScore. As was seen with the AU-ROC calculations, the latter three scoring functions performed very poorly when rescoring the poses from all five docking programs, while the former three functions generally performed well across the board. We note that in three instances, D-Score was able to detect and rank the active compounds significantly better than random, as demonstrated by the *p* values for DOCK docking of the ZINC and ACD decoy sets and FlexX docking of the ACD decoy set. These results follow very closely with the corresponding AU-ROC values. In all cases the native scoring functions were able to detect and rank the actives significantly better than random selection and ordering. Finally, when used to rescore docked poses, PMF-Score and F-Score each

**Table 3.4. Calculated SSLR Statistics with p Values for 5 Docking Programs, Native Score and Cscore Functions (Unrelaxed)**

| Docking Program / Validation Set | Native Score[a] | F-Score | PMF-Score | G-Score | D-Score | ChemScore |
|---|---|---|---|---|---|---|
| DOCK - ZINC | 212.3 (<.001) | 164.1 (<.001) | 145.8 (<.001) | 312.2 (.220) | 265.6 (.006) | 313.7 (.241) |
| DOCK - Schrödinger | 269.9 (.009) | 211.9 (<.001) | 201.7 (<.001) | 309.4 (.186) | 292.9 (.061) | 322.2 (.382) |
| DOCK - ACD | 183.2 (<.001) | 134.5 (<.001) | 182.6 (<.001) | 281.8 (.025) | 251.2 (.002) | 301.4 (.111) |
| FlexX - ZINC | F-Score | 244.1 (.001) | 151.3 (<.001) | 331.2 (.568) | 298.0 (.088) | 354.1 (.960) |
| FlexX - Schrödinger | F-Score | 270.3 (.009) | 204.5 (<.001) | 335.4 (.661) | 316.9 (.290) | 357.6 (.982) |
| FlexX - ACD | F-Score | 213.7 (<.001) | 190.3 (<.001) | 306.9 (.160) | 273.3 (.012) | 326.7 (.470) |
| Glide - ZINC | 131.4 (<.001) | 152.3 (<.001) | 148.6 (<.001) | 312.1 (.220) | 297.3 (.084) | 305.2 (.143) |
| Glide - Schrödinger | 125.9 (<.001) | 166.7 (<.001) | 192.7 (<.001) | 300.3 (.103) | 310.2 (.196) | 305.5 (.145) |
| Glide - ACD | 141.1 (<.001) | 141.3 (<.001) | 188.0 (<.001) | 312.0 (.219) | 279.2 (.021) | 279.5 (.021) |
| Surflex - ZINC | 112.1 (<.001) | 121.1 (<.001) | 140.0 (<.001) | 360.9 (.993) | 342.9 (.813) | 324.1 (.417) |
| Surflex - Schrödinger | 116.4 (<.001) | 134.4 (<.001) | 205.0 (<.001) | 370.0 (.999) | 363.2 (.997) | 346.7 (.878) |
| Surflex - ACD | 123.1 (<.001) | 132.9 (<.001) | 211.3 (<.001) | 353.5 (.956) | 333.0 (.607) | 325.4 (.443) |
| GOLD - ZINC | 259.6 (.004) | 169.5 (<.001) | 153.4 (<.001) | 349.6 (.916) | 344.5 (.842) | 358.6 (.986) |
| GOLD - Schrödinger | 254.6 (.002) | 203.0 (<.001) | 227.7 (<.001) | 352.3 (.945) | 365.2 (.999) | 360.1 (.991) |

a. SSLR statistics with *p* values <.05 are considered to have significant improvement over randomselection and ordering.

performed exceptionally well, matching their performance when gauged with the AU-ROC values.

In order to compare scoring functions to each other within docking program/decoy set pairs, *p* values were calculated to detect statistically significant differences in scoring function performance. Tables 3.5 through 3.7 show *p* value cross comparisons both for AU-ROC's and SSLR values for each of the three representative pairs mentioned above. These results are helpful in determining which, if any, of the top performing scoring functions significantly outperformed the other, or if there was no statistically significant difference. For example, in Table 3.5 the results indicate that between Glide Score, F-Score, and PMF-Score, there was no significant difference in their performance when judged by either AU-ROC or SSLR. Additionally, there is not a significant difference in the performance of D-Score, G-Score, and ChemScore when judged by either metric. In contrast, the data shown in Table 3.6 indicates that for Surflex docking of the Schrödinger decoy set, there was a significant difference between the performance of Surflex Score and PMF-Score that was detected by both metrics, with Surflex scoring significantly outperforming PMF-scoring. Additionally, it can be seen from Table 3.6 that a significant difference between PMF- and F-Score could not be detected from the AU-ROC values, but that a difference was detectable when comparing the two scoring functions with SSLR values. The ability of the SSLR value to detect a difference in performance of two scoring functions that was not detected by AU-ROC is also demonstrated in Table 3.7 when comparing PMF-Score and GOLD-Score, with GOLD-Score showing clear superiority over PMF-Score when judged by SSLR values. There are also instances where SSLR failed to detect a significant difference that was detectable by the AU-ROC method, as can be seen from the ChemScore/G-Score results in Table 3.7.

The results of a direct comparison of the native scoring functions to each other for each decoy set studied are given in Tables 3.8 through 3.10. It can be seen from the *p* values that Glide with its native Glide-Score and Surflex with its native Surflex-Score demonstrated a significant superiority over FlexX, GOLD, and DOCK with their own respective native scoring functions. Additionally, a direct comparison of Glide-Score and Surflex-Score shows that there is no significant difference between the results of the two scoring functions, both in terms of the AU-ROC and SSLR methods.

### 3.3.5 Post-Docking Relaxation

Several authors have recommended that, when rescoring poses with non-native scoring functions as reported here, the poses should first be optimized using the native scoring function before generating the score.[183,197] This procedure was not applied to the enrichment and AU-ROC data reported above, and it may explain the poor results observed with the D-Score, G-Score, and ChemScore algorithms. To investigate the effects of optimizing the ligand poses prior to rescoring, we applied the molecule relaxation function of Cscore to the docking output prior to rescoring with the five Cscore

**Table 3.5. Glide Docking of the ZINC Decoy Set**

| Docking Program / Validation Set | Glide Score | F-Score | PMF-Score | G-Score | D-Score | Chem Score |
|---|---|---|---|---|---|---|
| **Glide Score** | --- | .364 | .377 | <.001 | .001 | <.001 |
| **F-Score** | .674 | --- | .959 | <.001 | .006 | <.001 |
| **PMF-Score** | .717 | .957 | --- | <.001 | <.001 | <001 |
| **G-Score** | <.001 | <.001 | <.001 | --- | .281 | .712 |
| **D-Score** | <.001 | .011 | <.001 | .496 | --- | .212 |
| **ChemScore** | <.001 | <.001 | <.001 | .454 | .743 | --- |

**Table 3.6. Surflex Docking of the Schrödinger Decoy Set**

| Docking Program / Validation Set | Surflex Score | F-Score | PMF-Score | G-Score | D-Score | Chem Score |
|---|---|---|---|---|---|---|
| **Surflex Score** | --- | .267 | .018 | <.001 | <.001 | <.001 |
| **F-Score** | .319 | --- | .059 | <.001 | <.001 | <.001 |
| **PMF-Score** | <.001 | .003 | --- | <.001 | <.001 | <.001 |
| **G-Score** | <.001 | <.001 | <.001 | --- | .035 | <.001 |
| **D-Score** | <.001 | <.001 | <.001 | .042 | --- | .084 |
| **ChemScore** | <.001 | <.001 | <.001 | <.001 | .215 | --- |

**Table 3.7. GOLD Docking of the Schrödinger Decoy Set**

| Docking Program / Validation Set | GOLD Score | F-Score | PMF-Score | G-Score | D-Score | Chem Score |
|---|---|---|---|---|---|---|
| **GOLD Score** | --- | .373 | .297 | <.001 | <.001 | <.001 |
| **F-Score** | .112 | --- | .780 | <.001 | <.001 | <.001 |
| **PMF-Score** | .036 | .501 | --- | <.001 | <.001 | <.001 |
| **G-Score** | <.001 | .003 | <.001 | --- | .001 | .009 |
| **D-Score** | <.001 | .001 | <.001 | .038 | --- | .829 |
| **ChemScore** | <.001 | <.001 | <.001 | .251 | .916 | --- |

**Table 3.8. Native Scoring Functions with the ZINC Decoy Set**

| Docking Program / Validation Set | Grid Score | F-Score | Glide Score | Surflex Score | GOLD Score |
|---|---|---|---|---|---|
| **Grid Score** | --- | .822 | .078 | .042 | .015 |
| **F-Score** | .306 | --- | .019 | .007 | .659 |
| **Glide Score** | .085 | .030 | --- | .503 | .031 |
| **Surflex Score** | .050 | .002 | .676 | --- | .013 |
| **GOLD Score** | .236 | .647 | .018 | <.001 | --- |

**Table 3.9. Native Scoring Functions with the Schrödinger Decoy Set**

| Docking Program / Validation Set | Grid Score | F-Score | Glide Score | Surflex Score | GOLD Score |
|---|---|---|---|---|---|
| **Grid Score** | --- | .808 | .004 | .004 | .871 |
| **F-Score** | .988 | --- | .001 | .001 | .689 |
| **Glide Score** | <.001 | <.001 | --- | .703 | .003 |
| **Surflex Score** | .003 | .002 | .829 | --- | .001 |
| **GOLD Score** | .603 | .714 | <.001 | <.001 | --- |

**Table 3.10. Native Scoring Functions with the ACD (Bissantz) Decoy Set**

| Docking Program / Validation Set | Grid Score | F-Score | Glide Score | Surflex Score |
|---|---|---|---|---|
| **Grid Score** | --- | .866 | .146 | .120 |
| **F-Score** | .277 | --- | .043 | .048 |
| **Glide Score** | .317 | .080 | --- | .888 |
| **Surflex Score** | .160 | .017 | .707 | --- |

scoring functions. This relaxation function uses the Tripos Force Field to perform a 100 iteration torsional minimization of the docked ligand. Figure 3.12 and Figure 3.13 show the effects of this relaxation procedure on the rescored AU-ROC values for the poses scoring functions. This relaxation function uses the Tripos Force Field to perform a 100 iteration torsional minimization of the docked ligand. Figure 3.12 and Figure 3.13 show the effects of this relaxation procedure on the rescored AU-ROC values for the poses generated by Surflex docking of the ZINC decoy set. There was little change in the calculated AU-ROC values for D-Score, some improvement for G-Score, and significantly decreased AU-ROC values for the F-Score, ChemScore, and PMF-Score functions. Similar results were obtained for all the docking programs and decoy sets investigated in this study (data not shown).

### 3.3.6 Consensus Scoring

Consensus scoring has received mixed reviews in recent validation studies, with some authors reporting enhanced enrichment over single scoring functions[85,86,92] and others reporting little to no improvement.[184,191] To further investigate this in the DHPS system, we used the Cscore module of Sybyl 7.3 to generate consensus scores from the five Cscore functions. We used the default settings, and investigated consensus score values generated from both unrelaxed and relaxed scores. A score of 0 through 5 is generated for each ligand pose depending on the number of "good" scores received from each of the five C-score functions. Table 3.11 gives the results of consensus scoring on enrichment (by calculated AU-ROC) for each of the five docking programs. Only the data from the ZINC decoy set are shown in the table, but the results were similar for the other two decoy sets. Table 3.11 gives the results from the unrelaxed and relaxed poses for comparison. Ideally, the majority of the known active compounds should give a high Cscore value of 4 or 5, while the majority of the decoy compounds should have low Cscore values. However, consensus scoring resulted in only a modest enrichment, and it failed to significantly improve the enrichment results obtained when scoring with single scoring functions. We saw no significant difference in the results when Cscore calculations were performed on the unrelaxed poses over the relaxed poses. The best results (an AU-ROC value of 0.891) were seen with consensus scores generated from unrelaxed poses from the Glide docking.

## 3.4 Discussion

Our high resolution crystallographic studies of DHPS from *Bacillus anthracis* that includes substrate and inhibitor complexes has provided us with the opportunity of using virtual screening methods to identify novel inhibitory compounds that specifically dock into the well characterized binding determinants of the pterin pocket. However, an to identify which compounds should be further pursued by in-depth biochemical, kinetic and structural studies. We have therefore performed a thorough investigation of docking

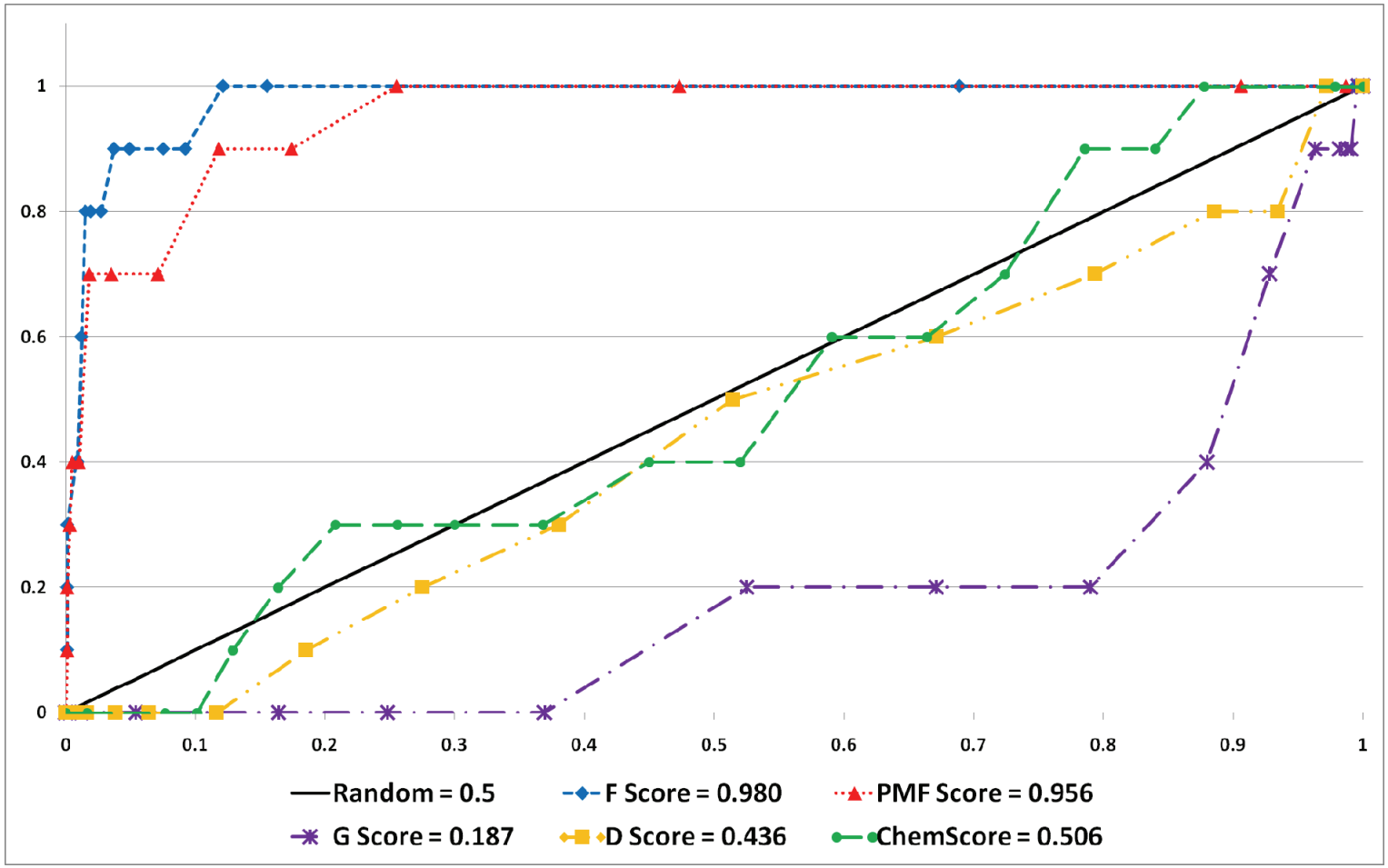



**Figure 3.12. Effect of Molecule Relaxation of Docking Output Prior to Rescoring with Cscore Functions (Unrelaxed)**

Surflex docking of the ZINC validation set is shown unrelaxed.

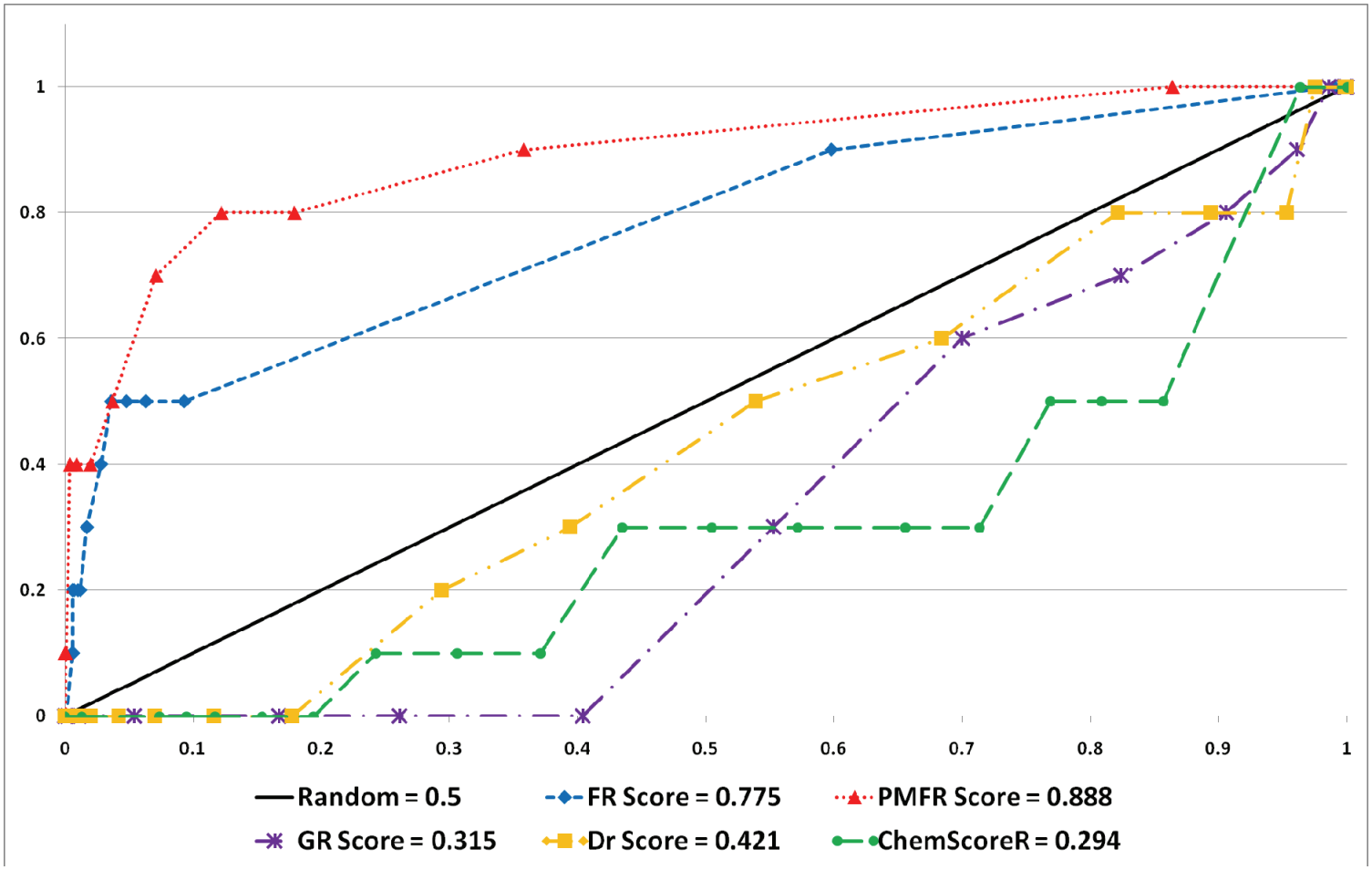



**Figure 3.13. Effect of Molecule Relaxation of Docking Output Prior to Rescoring with Cscore Functions (Relaxed)**

Surflex docking of the ZINC validation set is shown relaxed.

**Table 3.11. Cscore AU-ROC Results for Docking of ZINC Decoy Set**

| Docking Program / Validation Set | Unrelaxed | Relaxed |
|---|---|---|
| **DOCK** | .730 | .734 |
| **FlexX** | .722 | .679 |
| **Glide** | .891 | .857 |
| **GOLD** | .716 | .658 |
| **Surflex** | .622 | .554 |

and scoring methodologies to identify which combination would be expected to yield the best results when applied to this particular pocket in this particular enzyme. As described in the introduction, we sought answers to eight specific questions and have successfully provided key insights into each of them.

We first investigated pose selection and noted the overall good performance of all five docking programs. Each program was able to generate a successful pose (RMSD less than 1.5 Å), and four of the five native scoring functions were able to rank a successful pose. Additionally, when the poses were rescored with the five Cscore scoring functions, each one performed reasonably well. The majority of the docking and scoring functions were able to generate and rank successful poses, and we therefore conclude that this method of evaluating docking/scoring combinations is useful for eliminating poorly performing combinations but not for selecting the optimal combination.

We then addressed the question of how two commonly-used metrics, enrichment calculations at a given percent of decoy set screened (1% and 2%) and areas under receiver-operating characteristic curves (AU-ROC), compare when used for validation. Although both metrics were generated using the same data, it was easier to note a difference in performance when analyzing the enrichment values. Using the AU- ROCs, we classified combinations as performing either well, moderately well, or poor, but within each classification, it was difficult to determine the best docking/scoring combination. Similar to the pose selection study, the AU- ROCs were most useful for eliminating poorly performing docking/scoring combinations rather than selecting the top performing combination. In contrast, the enrichment calculations which reward early detection of active compounds appear to be more successful in distinguishing the top performing docking/scoring combinations for use against a specific target, based on our results.

We next sought to answer the question of how important is the selection of decoy compounds for use in enrichment studies. A recent study stressed the importance of selecting decoy set compounds that closely match the active compounds in terms of physico-chemical properties in order to avoid artificial enrichment.[218] We selected three decoy sets that had previously been used to validate docking programs against a wide

variety of enzyme targets. Each of the three decoy sets has slightly different characteristics, and differs in physico-chemical properties from the active compounds to varying degrees. We compared the enrichment and AU-ROC results across decoy sets, and although there were detectable differences when comparing enrichment calculations at 1 and 2%, we were unable to correlate this trend with the degree of difference in physico-chemical properties of the active compounds from the decoy sets. Significantly for our purposes, when comparing the AU-ROC calculations across decoy sets, we did not detect a significant trend either favoring or disfavoring one decoy set over another, and when a docking/scoring combination performed well, it generally did so against all three decoy sets and vice versa. However, our results are not necessarily inconsistent with the previous study where a trend was observed.[218] More likely, while our active compounds differed significantly in some physicochemical properties from the 3 decoy sets we selected, the decoy sets themselves did not differ enough between each other to make a clear distinction in performance. This can be seen in Table 3.12, which shows the average molecular volume, atom count, cLogP, # of H-bond donors (HD), # of H-bond acceptors, and number of rotatable bonds (RB) for the active set and the three decoy sets. It can be seen that the active compounds tend to be smaller and more hydrophilic than the decoy compounds, but the decoy sets themselves are very similar.

The question of how to deal with docking failures was also specifically addressed because this issue has received little attention in previous studies. In our study, the docking programs were frequently unable to return poses for some decoy compounds, and this led to a problem in directly comparing the programs. For example, the program Glide in combination with the Schrödinger decoy set returned 607 successful poses while the Surflex program returned the full quota of 1010 poses (1000 decoys plus 10 actives). In the calculation of enrichment, we believe that it would have been an unfair penalty on programs that failed to dock decoy compounds had we selected % compounds docked rather than the % of the total number of compounds, and similarly in the calculation of AU-ROC and SSLR. We therefore used the total number of compounds (decoy + active) to calculate enrichment at 1 and 2%, and assigned the worst reported score to all docking failures when calculating the ROC plots, AU-ROC's, and SSLR values. We recognize that this method may over-compensate because the failure of a program to dock an inactive compound may actually reflect superior performance. Thus, in the event that the performance of two docking/scoring

**Table 3.12. Active and Decoy Compounds Average Characteristics**

| Set | Volume | Atom Count | cLogP | HD | HA | RB |
|---|---|---|---|---|---|---|
| Actives | 211 | 37 | 0.29 | 3 | 7 | 5 |
| ACD | 287 | 42 | 3.27 | 2 | 4 | 5 |
| Schrödinger | 341 | 50 | 3.77 | 2 | 6 | 6 |
| ZINC | 310 | 42 | 3.02 | 1 | 4 | 5 |

combinations are indistinguishable, we believe it is reasonable to use the number of docking failures for inactive compounds as a means for selecting one over the other.

We next addressed the abilities of post-docking relaxation and consensus scoring to improve enrichment results by evaluating their effects on our AU-ROC metrics. Cole and co-workers have stressed the importance of using scoring functions to optimize docking output prior to rescoring with that function[197] and we attributed the poor performance of the D-Score, G-Score, and ChemScore functions to this deficiency in our analyses. To test this, we performed molecule relaxation using a function that is available in the C-Score module of Sybyl (Tripos) prior to rescoring, and compared these results with the unrelaxed scores. The scoring results following relaxation were typically worse in terms of AU-ROC, and this may be due to the function's use of the Tripos Force Field rather than the scoring functions themselves. This is consistent with the findings of Cole and coworkers because the notable exception was the improved performance of G-Score that actually uses the Tripos Force Field parameters. Consensus scoring failed to improve upon the results we were able to obtain with single function scoring, and we believe that this can also be attributed the fact that the Cscore functions were not optimized with respect to the functions themselves. We conclude that, when rescoring with non-native scoring functions, it is very important to optimize with respect to that scoring function.

The known inhibitory constants of the active compounds seeded into the decoy sets represents important information that can be used to further evaluate the performance of docking and scoring combinations. Thus, ideally, the active compounds should not only be identified early but also in the correct order according to inhibition constants. In this study we have introduced a new method for interpreting enrichment study results that simultaneously rewards early detection of active compounds and correct ordering, the 'sum of the sum of log rank' or SSLR. Although several methods have been reported that specifically reward early detection[203,219], we believe that this is the first method that takes this approach. The SSLR method was developed to help us distinguish between the top performing docking and scoring combinations that were statistically indistinguishable using traditional AU-ROC methods. In the three representative examples given above, the SSLR method was able to distinguish between scoring functions in two cases where the differences in AU-ROC were not significant but, in general, the SSLR values closely correlated to the AU-ROC results in terms of statistical significance. However, it is very straightforward to apply the SSLR method when relevant data are available, and we consider this a valuable method with potentially great utility for future virtual screening studies.

The ultimate goal of this study was to determine which of the docking and scoring combinations evaluated would be expected to yield the best results in terms of enrichment when used against the pterin binding site of DHPS in a large scale, virtual screening study. We noted the excellent performance of the native scoring functions when used with each of their respective docking programs in our enrichment studies.

We also noted the poor performance of the Cscore scoring functions when used to rescore docking output, and explained this by our inability to optimize the poses with respect to the scoring functions themselves. While this may explain the poor performance of G-Score, D-Score, and ChemScore, it does not explain the good to excellent performance of F-Score and PMF-Score. We believe that the nature of the pterin binding site may in part explain this observed phenomenon. Ligand binding into the pterin binding site not only involves van der Waals packing interactions within the tight pocket but also polar hydrogen bonding and ionic interactions.[154] Additionally, as can be seen from Figure 3.4, there is a clear preference for planar, aromatic compounds that can accommodate π-stacking with the side chain of Arg254. Our results are consistent with those of Bissantz and co-workers who found that FlexX scores and PMF scores performed better against polar active sites, while DOCK scores were more reliable against non-polar active sites.[196] There is also an explicit aromatic stacking term used in F-Score, unique to this scoring function, which may have also contributed to its good performance.

## 3.5 Summary

In order to select the best performing docking/scoring combination for virtual screening studies against the DHPS pterin binding site, we employed several validation methods. Pose selection studies using a co-crystal structure with a known pterin-site inhibitor bound were useful in identifying docking/scoring combinations that performed poorly but were less helpful in selecting a top performing combination. Similarly, the AU-ROC values were also less helpful at selecting a specific top-performing docking/scoring combination, but clearly identified poorly performing combinations. However, enrichment calculations at 1 and 2% percent of the decoy set screened proved very useful in identifying two top performing docking/scoring combinations, Glide with Glide Score and Surflex with Surflex Score. Finally, we have developed a new metric that can be used as a validation method that we term SSLR. The SSLR statistic not only takes into account early detection of active compounds from decoy sets, but also rewards for correctly ordering the active compounds by their known inhibitory constants. We found that the results of the SSLR tests closely matched the AU-ROC results and in several cases were able to help us distinguish between docking/scoring combinations for which there was not a statistically significant difference using the latter method.

We investigated three separate decoy sets and found a dependence on the decoy set used when calculating enrichment at 1% and 2%, with the ZINC decoy set yielding the highest enrichment values. This dependence was not seen when comparing AU-ROC's from ROC plots, which were generally comparable across validation sets. Our investigations also showed that relaxation of the poses prior to rescoring with the Cscore functions using the relaxation function of the Cscore module implemented in Sybyl 7.3 did not overall improve enrichment and in some cases was actually detrimental. We believe this is due to the fact that the Cscore relaxation function does not use the scoring function to minimize the poses, but instead uses a different force

field. No improvement over the best results seen with single scoring functions was observed when applying consensus scoring, with either the relaxed or non-relaxed poses. Again, we postulate that this is due to the fact that the Consensus scoring functions were not optimized with respect to each function prior to scoring.

We demonstrate considerable variability when using these various validation methods and identify clear winners. Indeed, without these analyses, it would be virtually impossible to successfully use virtual screening in our studies. Based upon the results from the enrichment studies, AU-ROC and SSLR calculations, we found that, of the docking programs and scoring functions we evaluated, the most appropriate combination for use in high-throughput virtual screening against DHPS would be Glide with the native Glide Score function or Surflex with the native Surflex Score function.

# CHAPTER 4. HIGH-THROUGHPUT VIRTUAL SCREENING AGAINST DHPS

## 4.1 Introduction

The work presented in this chapter will complete my discussion of the DHPS virtual screening project that began in Chapter 2 and continued in Chapter 3. Chapter 2 discussed the crystal structure of *B. anthracis* DHPS and the molecular simulation studies that were performed to investigate the structure and function of the flexible loop regions surrounding the DHPS active site as well as to develop a complete screening structure for high-throughput virtual screening studies. In Chapter 3, I presented the results of a large-scale, validation study that was performed to select the best docking and scoring combination for use in high-throughput docking studies against the pterin binding site of DHPS. This chapter will introduce the methods that were utilized to screen several million compounds against the pterin site and the results of those investigations.

### 4.1.1 The DHPS Pterin Binding Site

We have solved several crystal structures of the *B. anthracis* DHPS enzyme with both substrate and product analogs as well as an inhibitor bound.[154] These structures have shown that the active site can be separated into sub-sites for the binding of the pterin substrate and the *p*ABA substrate, as shown in Figure 4.1. The sulfonamide agents, as previously mentioned, bind to the *p*ABA sub-site and inhibit product formation or combine with the pterin substrate to form “dead-end” products. Mutations that confer sulfonamide resistance have been mapped to the DHPS enzyme and fall near the *p*ABA binding site, as shown in Figure 4.2. Theoretically, agents that inhibit the DHPS enzyme by binding to the pterin sub-site would be able to bypass the resistance mutations that have rendered the sulfonamide agents less useful for the treatment of infection. Table 4.1 lists the key binding residues in the pterin site and their corresponding residues in several common pathogenic bacterial organisms. It can be seen that that there is a high degree of conservation among the key binding residues between these different species. This implies that inhibitors of the pterin binding site of DHPS may have low species specificity and could result in antibacterial agents with a broad spectrum of activity against many Gram positive and Gram negative bacterial species.

### 4.1.2 Virtual Screening against DHPS

As discussed in Chapter 1, virtual screening has been shown to be complementary to high-throughput screening as a method to identify lead compounds in a drug design project.[11] Structure-based virtual screening involves the use of a 3D structure of the drug target, usually by X-ray crystallography or NMR studies, and

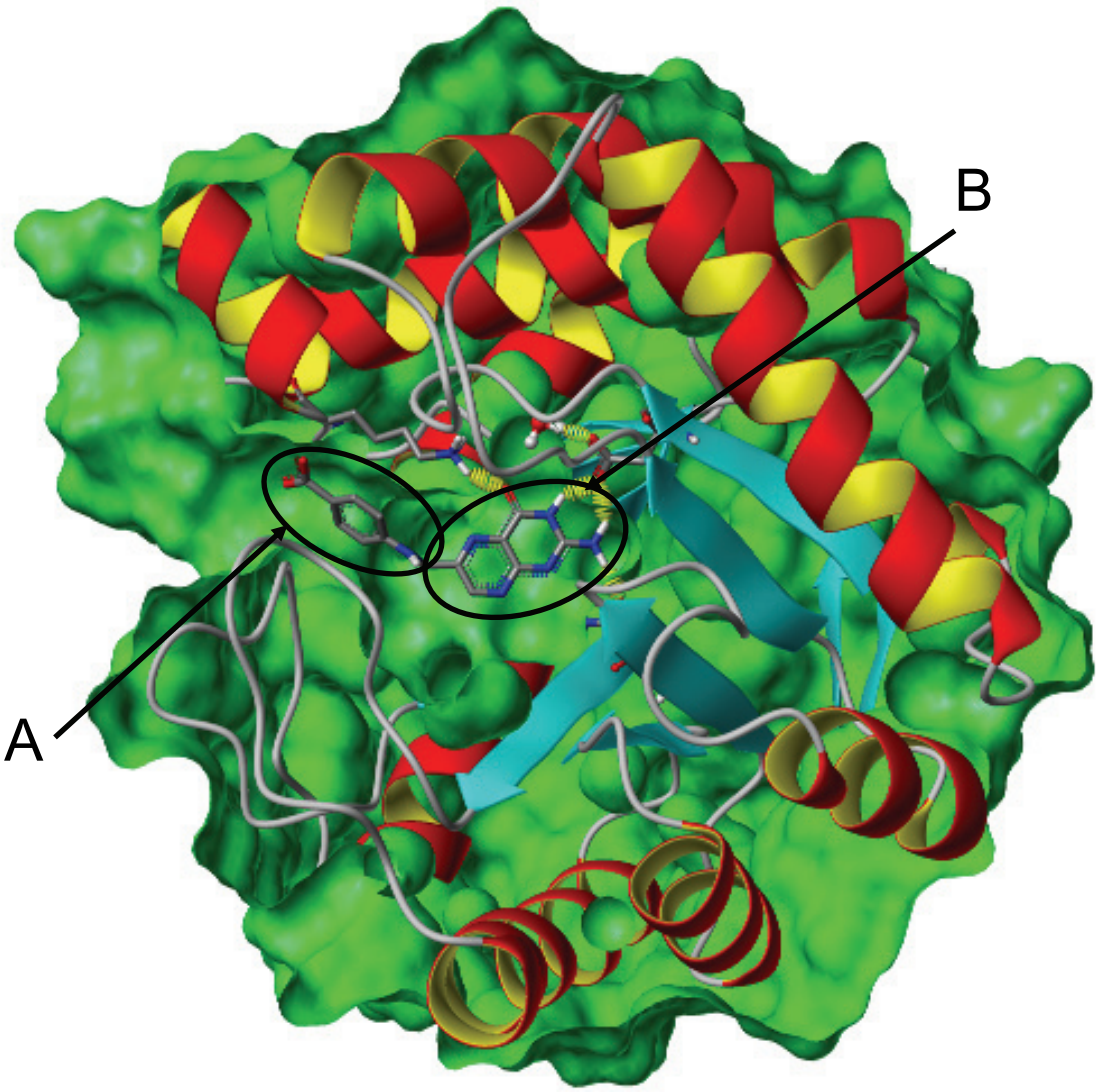



**Figure 4.1. DHPS with Pteroate Product Analog Shown Bound**

DHPS with *p*ABA and Pterin binding sites indicated using a pteroate product analog. A. *p*ABA binding site falls near the solvent exposed surface, enclosed by flexible loop regions. B. Pterin binding site deep within enzyme in a highly conserved pocket.

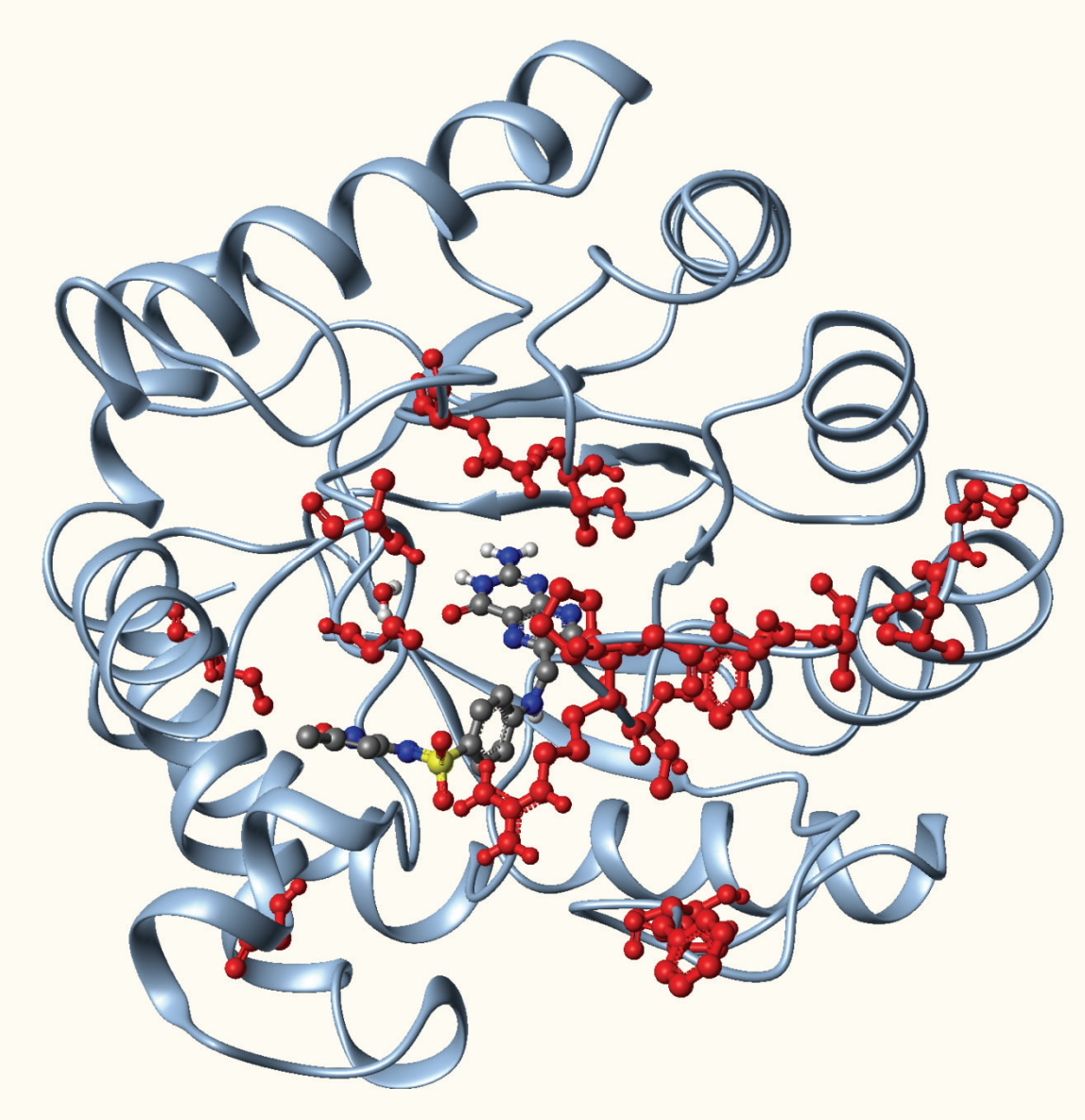


**Figure 4.2. The DHPS/Pterin-SMX Structure with Sulfonamide Resistance Conferring Mutation Sites Indicated**

Residues that confer resistance to sulfonamide antibiotics in several species (see discussion in Chapter 2) have been mapped to the *B. anthracis* DHPS structure and are shown here in red. The mutation sites predominately fall on flexible loop regions near the *p*ABA (sulfonamide) binding pocket.

**Table 4.1. DHPS Pterin Binding Site Residues**

| *B. anthracis* | Interaction Type | *E. coli* | *S. aureus* | *M. tuberculosis* | *S. pneumoniae* | *P. aeruginosa* |
|---|---|---|---|---|---|---|
| Thr67[a] | n/a | Thr62 | Thr51 | Ser53 | Thr57 | Ser49 |
| Arg68[a] | n/a | Arg63 | Arg52 | Arg54 | Arg58 | His50 |
| Asp101 | vDw? no direct | Asp96 | Asp84 | Asp86 | Asp91 | Asp82 |
| Asn120 | H-Acceptor | Asn115 | Asn103 | Asn105 | Asn110 | Asn101 |
| Ile122 | vDw | Ile117 | Gln105 | Val107 | Ile112 | Ile103 |
| Ile143 | vDw | Cys137 | Val126 | Val128 | Val133 | Val123 |
| Met145 | Pi Electronic | Met139 | Met128 | Met130 | Met135 | Met125 |
| Asp184 | H-Acceptor | Asp185 | Asp167 | Asp177 | Asp201 | Asp173 |
| Phe189 | Pi Electronic | Phe190 | Phe172 | Phe182 | Phe206 | Phe178 |
| Leu214 | vDw | Leu215 | Leu197 | Leu207 | Phe231 | Leu206 |
| Gly216 | no direct | Gly217 | Ala199 | Gly209 | Gly233 | Ser208 |
| Lys220 | H-donor | Lys221 | Lys203 | Lys213 | Lys237 | Lys212 |
| Arg254 | Pi Electronic | Arg255 | Arg239 | Arg253 | Arg282 | Arg246 |

[a.]Residues which fall on mobile loop elements and have been modeled in to place. Their position is uncertain or unknown in several species.

Residues differing from *B. anthracis* target are colored in red.

molecular docking experiments in which corporate or commercial libraries are docked and scored in the target's active site to identify compounds with binding affinity. Because commercial libraries of screening compounds can be quite large, on the order of several million compounds for the larger libraries, the computational expense of docking the entire library can be significant, even with the most efficient docking programs. For this reason, often specific constraints are applied to screening libraries prior to docking which may include 1D physicochemical property filters like the commonly used "Rule of Five" proposed by Lipinski and co-workers, which describes filters for oral, drug-like compounds.[23] Another type of constraint that can be applied to limit the size of the screening library to be docked is the pharmacophore constraint. This requires specific knowledge of key binding interactions that should be conserved for successful inhibitor binding, usually obtained from co-crystal structures with ligand or inhibitor bound. Using this method, screening libraries can be filtered to only include compounds with specific numbers and locations of key binding elements such as hydrogen bond donors or acceptors, aromatic rings, lipophilic groups, etc. Finally, fragment constraints can be applied using the "Rule of Three", or a variation thereof, which filters for smaller, more fragment-like compounds.[130]

The advantages of screening fragments over drug-like compounds were discussed in Chapter 1; most notable is the likelihood that the lead optimization process will result in a drug-like compound which has a greater chance of having good oral bioavailability and favorable ADME properties. Additionally, a much smaller number of compounds are generally needed for successful fragment-based screening, usually on the order of a hundred to a few thousand. This is because compounds of lower complexity have a greater chance of matching the target receptor site. Fragment-based screening is not without its disadvantages. Because of the lower molecular weight and complexity of the fragment compounds, they are expected to be less potent than a drug-like compound. This means that specialized screening methods need to be employed to identify hit compounds. Several methods have been used with success, including high concentration screening[133] (up to mM concentration), X-ray crystallography or NMR screening,[135] affinity detection by mass spectrometry,[136] and surface plasmon resonance,[137] and ITC.[138] It should be noted that although the fragment hits will show a much lower potency, often high micromolar to low millimolar, in terms of binding efficiency (binding affinity normalized by molecular weight or heavy atom count), they are often on par with or exceed the efficiency of drug like compounds.[139]

### 4.1.3 Research Goals and Design

Utilizing an X-ray crystal structure from *B. anthracis* DHPS with an inhibitor bound into the pterin site, we performed several large-scale, high-throughput virtual screens using the docking methods validated in Chapter 3. The virtual screening followed two strategies that we implemented in two successive rounds of high-throughput docking. In the first round, a pharmacophore filter based upon the key pterin site binding elements was applied to compounds from the ZINC databases.[220]

Compounds which passed the filter and a molecular weight cut-off were subsequently docked into the DHPS pterin site using our validated docking method, as described below. In the second round, we forewent the pharmacophore filter and applied a molecular weight and rotatable bond constraint to a subset of the commercial libraries used in the first round. The compounds passing this constraint were then docked into the pterin site using the validated docking method.

The pharmacophore filter was removed from the second round of docking for several reasons. First, we hoped that by removing the pharmacophore pre-docking step we could identify compounds that were unlike the pterin substrate in appearance and physical property. The goal was to discover novel scaffolds with improved solubility that could be taken into subsequent lead optimization trials. Second, removing the pharmacophore filter provided us with an opportunity to investigate the use of pharmacophore pre-processing of screening libraries versus simple high-throughput docking, in terms of hit rates, quality of hits, and computational expense. Finally, the second round of docking allowed us to compare the two top performing docking programs from our validation study (presented in Chapter 3) in actual screening studies against the target enzyme.

Hit compounds from both rounds of virtual screening were selected for testing in our enzyme assay. Fragment hits which displayed greater than 30% inhibition of DHPS activity in our assay were selected for investigation in crystallography trials. The results of the two rounds of virtual screening with hit compound inhibitory activities are presented herein.

## 4.2 Computational and Experimental Methods

### 4.2.1 The DHPS Screening Structure

A crystal structure of the *B. anthracis* DHPS enzyme with an inhibitor known to bind the pterin site, AMPPD (Figure 4.3), has been solved and was used for all the molecular docking performed in this study. Flexible loops 1 and 2 are highly mobile elements. It is believed that the loops close over the active site after PtPP binding, forming the *p*ABA binding pocket and facilitating enzymatic catalysis.[153] As can be seen from Figure 4.2, the majority of sulfonamide resistance mutation sites fall on loops 1 or 2. Although visible in our crystal structure, the position of loop 1 is believed to be a crystallization artefact due to contact with a neighbouring monomer, rather than occupying a functional position (Figure 4.4, left). The positions of several residues, 66-74, in loop 2 are disordered and not visible in this crystal structure. The missing or inaccurately positioned residues from loops 1 and 2 were investigated by homology modelling and extensive molecular dynamics simulations as discussed in Chapter 2. The initial positions for loops 1 and 2 for these docking studies were taken from the average structure obtained in our MD simulation series 2-17, which was performed using

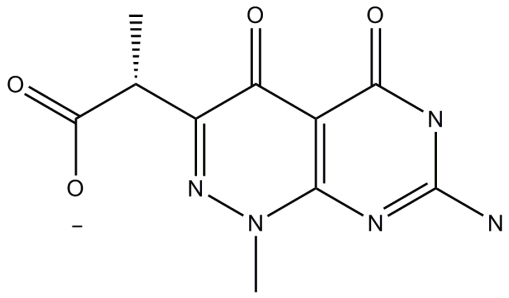



**Figure 4.3. AMPPD Structure**

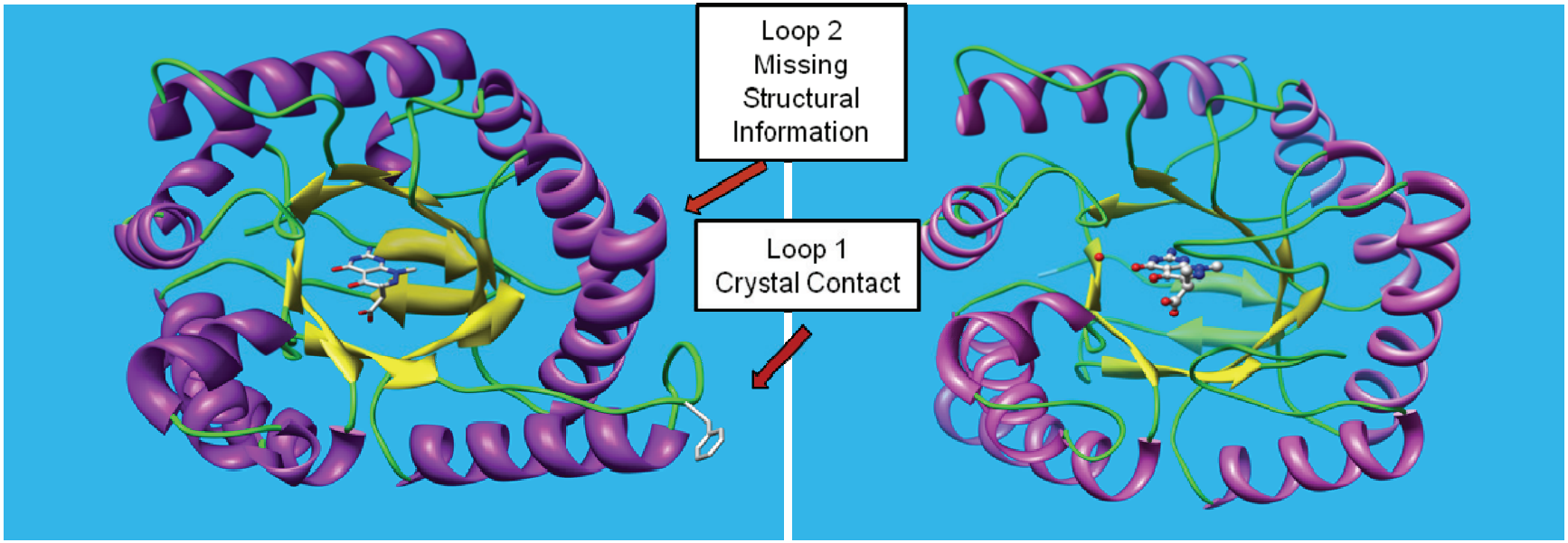



**Figure 4.4. *B. anthracis* DHPS Before and After Flexible Loop Placement**

*B. anthracis* DHPS Structure is shown before and after preparation for docking. The positions of mobile loops 1 and 2 were homology modeled from the *M. tuberculosis* and *E. coli* crystal structures and minimized by molecular dynamics and energy minimization methods, as discussed in Chapter 2 methods section.

our crystal structure with the most residues from loop 2 visible to date. Loops 1 and 2 fall near the *p*ABA binding pocket during catalysis and are not believed to play a large role in pterin substrate binding. To prepare the enzyme for docking, hydrogens were added and AMBER FF99 charges were calculated for the protein. A water molecule located near the pterin binding pocket is conserved in all the DHPS structures published thus far is believed to be structurally required. Charges were loaded to the water and it was left in the active site for all docking runs performed. Hydrogen positions were refined by performing a 1000 iteration minimization with heavy atoms constrained using the Powell method with initial Simplex optimization.[221] Figure 4.4 shows the docking structure before and after placement of the flexible loops.

### 4.2.2 The Docking Protocol

The docking validation study reported in Chapter 3 concluded that for high-throughput docking studies involving DHPS, Surflex-Dock and Glide-Dock perform exceptionally well. In this study, 2 rounds of high-throughput docking were performed, one using each docking function. The first round involved docking of the ZINC version 6 (2006) databases after pre-filtering with molecular weight and pharmacophore constraints using the UNITY program available in Sybyl 7.3. The ZINC databases contain over 4.6 million compounds, and include multiple tautomers and protonation states for the screening compounds. In order to filter for fragment-like compounds and decrease the number of compounds requiring docking to a more manageable number, we used a molecular weight filter of 350 Daltons for the first round of screening and also employed pharmacophore filters as described below. Compounds meeting the pharmacophore criteria were then docked and scored in the DHPS pterin binding site with the Surflex docking tool using the same docking methods described below. The top 2% of the Surflex-Score ranked compounds were selected for testing in the DHPS enzyme assay.

Round 2 of the high-throughput virtual screening forewent the pharmacophore filters in an attempt to identify scaffold compounds that were not “pterin-like”. Several commercial vendors were identified based upon the ease of acquisition of their compounds (from Round 1) as well as the availability of their screening sets in an easily obtainable format for screening. The compound screening sets were obtained, filtered by modified “Rule of Three” criteria, and docked using the Glide-Dock program of Schrodinger, Inc. The highest scoring compounds were selected by score and diversity for enzyme assay and crystallography using the methods described below.

### 4.2.3 UNITY Database Preparation

The screening compounds used in Round 1 were downloaded in .sdf format from the vendors located in the ZINC version 6 libraries.[220] The libraries contained over 4.6 million compounds including protonation variants and tautomers for the medium pH

range of 5.75 to 8.25 in 26 different vendor sets. Additionally, these libraries have been pre-filtered to remove reactive and cytotoxic compounds as discussed in Chapter 1 (the ZINC database filtering rules can be found in Appendix D). The .sdf files were converted to UNITY databases for pharmacophore screening using the UNITY program available in the Sybyl 7.3 molecular modeling suite of Tripos, Inc.[222,223] During preparation, 2D and Macro fingerprints were created using Unity's default settings. The Concord program was used to generate 3D coordinates, when necessary.[224] Default values were accepted for all other UNITY database preparation settings.

### 4.2.4 Pharmacophore Filtering

The pharmacophore filters utilized in the first round of virtual screening are shown in Figure 4.5. They were created and applied to the ZINC screening sets using the UNITY program. A surface volume constraint (Figure 4.5, top) was created using all pterin site residues falling within 8 Å of the AMPPD ligand with a VdW tolerance of 1 Å. Macros were created for 1 donor and 4 acceptor positions based upon the H-bonding patterns seen with the AMPPD ligand as well as the pterin substrate (Figure 4.5, bottom). A spatial tolerance of 0.3 Å was used for each macro and 2 partial match constraints were applied to loosen the filter and remove false carboxylate and ester hits.

The UNITY databases created from the ZINC screening libraries were screened against this pharmacophore model using a Flex search with modified "Rule of Three" search options specified. The maximum molecular weight was 350 Daltons and the maximum number of rotatable bonds was five. Flex ring search option was also enabled. All other settings retained their default values.

### 4.2.5 Docking Library Preparation

For the first round of virtual screening, hitlists from the pharmacophore filtering stage were merged to eliminate duplicate compounds and then the converted to a multi-mol2 file for docking. Charges were loaded to the compounds using the Gasteiger-Huckel method.[105] The compounds were docked using the Surflex docking function and scored with the Surflex scoring function as described below.[65] The experimental docking methods for both Surflex and Glide docking (virtual screening round 2) were the same as used in our validation study discussed in Chapter 3.

Compounds used in the second round of virtual screening were obtained directly from the chemical suppliers as .sdf files of their most updated collections, rather than downloaded as sets from the ZINC site. The following chemical suppliers were used due to the ease of obtaining compounds, reliability (in terms of compound purity), or ease in obtaining their screening library in electronic format: ASDI, ChemBridge, ChemDiv, InterBioscreen, Key Organics, Life Chemicals, MayBridge Ro3 screening set, Maybridge complete set, Nanosyn, Peakdale, Pharmeks, Ryan Scientific, Sigma Aldrich,

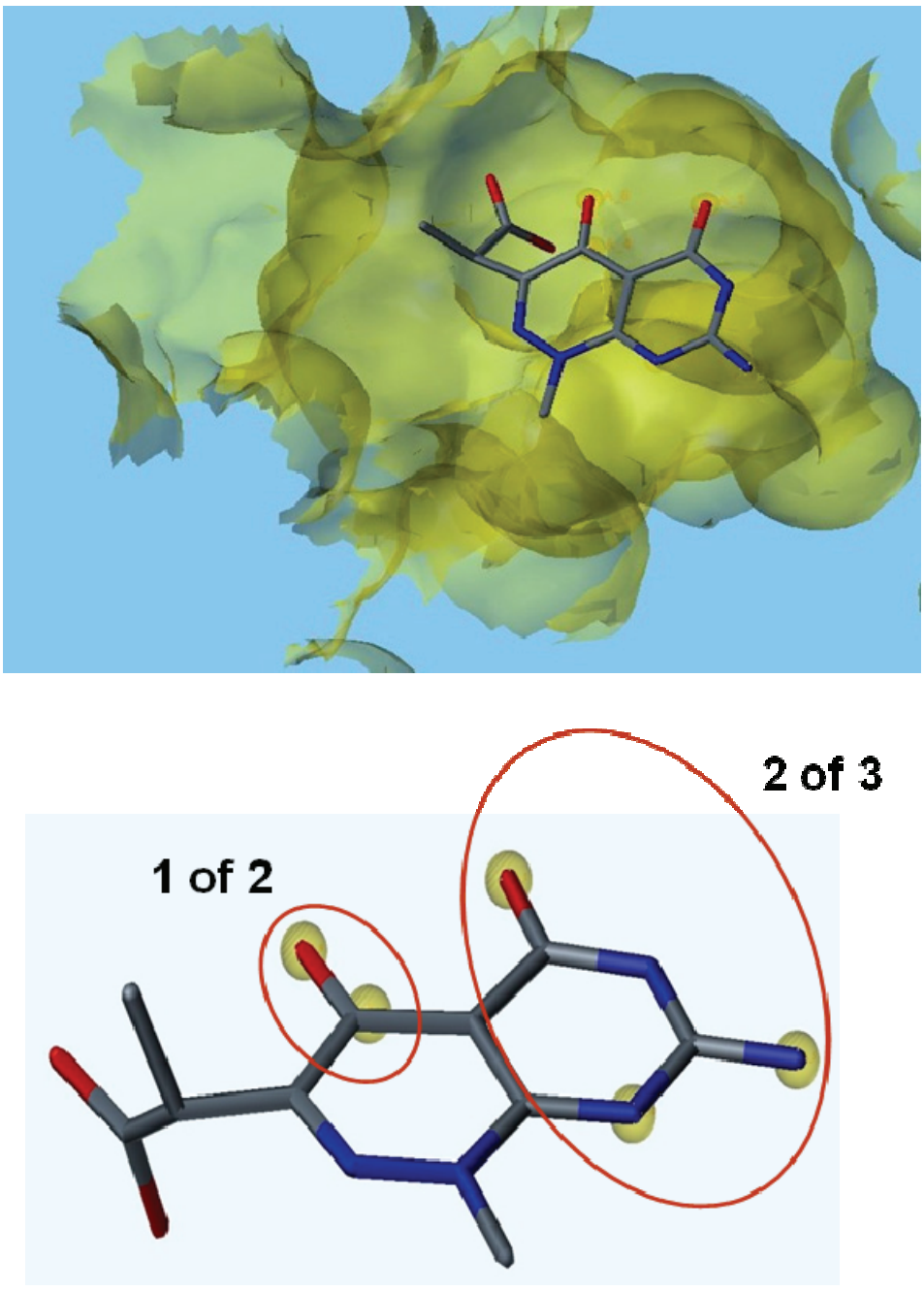



**Figure 4.5. UNITY Pharmacophore Filters Applied to DHPS**

Top: Active site surface
Bottom: Hydrogen Bond donor and acceptor macro filters.

Specs, SynChem, Synthonix, and TimTec. The .sdf files were converted to .mae files for Glide docking using the LigPrep program of Schrodinger, Inc. Fragment filters included rotatable bond count of 5 and molecular weight of 300 Daltons. Ionization states were built using Epik for a target pH range of 7.2 to 7.6. The desalter function was employed to remove waters and counter ions and tautomers were generated. Stereoisomerism was retained if specified and varied when not specified. Low energy ring conformations were obtained, hydrogens added and 3D conformations generated using the OPLS 2005 force field.

### 4.2.6 Docking, Scoring, and Processing

Surflex docking in the first round utilized the multi-mol2 files generated as described above and a protomol generated using a threshold of 0.50 and bloat of zero (default values). These settings are the same as those used in the docking validation study discussed in Chapter 3. An active site water (Figure 4.4, right) was retained for all docking runs. The ring flexibility function was enabled; all other docking settings retained their default values. Compounds docked with Surflex were scored with the native Surflex scoring function; the Cscore option was disabled. The top 2% of the Surflex scored compounds were selected for procurement and testing in the enzyme assay described below.

Glide docking of the fragment compounds in the second round of virtual screening utilized the .mae files generated for each supplier as described above. The Glide receptor grid was generated using the *B. anthracis* DHPS structure described above and an active site defined by an 8 angstrom box around the AMPPD ligand. Default van der Waals radius scaling settings were employed for generation of the receptor grid. No docking constraints were defined. Compounds were docked using the standard precision setting with the flexible docking option enabled. All other Glide docking settings used default values. These settings are the same as those utilized during the validation of this docking method against the DHPS target. The top 1% of docked compounds from each supplier were selected by Glide score and merged into a single file, resulting in a set of 2845 compounds. Because the assay employed in this study is not a high-throughput assay (see assay methods below), it was not feasible to test all 2845 compounds in the assay within a reasonable amount of time. Therefore a diversity filter was applied to the high-scoring compounds using the Selector program available in the Sybyl 7.3 molecular modeling suite. The compounds were saved as a multi-mol2 file and imported as a Sybyl Molecular Spreadsheet. The diversity metrics employed were 2D fingerprints and Atom Pairs with equal weighting. The hierarchical clustering method was used to generate 54 clusters. The highest scoring compound in each cluster was then selected for testing in the enzyme assay.

## 4.3 High-Throughput Virtual Screening: Results and Discussion

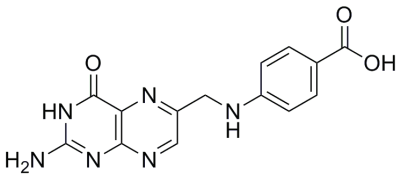
### 4.3.1 Pharmacophore-Based Virtual Screen

5,093 compounds from the ZINC screening libraries matched the pharmacophore requirements of the first round of virtual screening. When the UNITY hitlists were merged, the total number of unique compounds was 3104, indicating a large degree of redundancy in the ZINC databases. All 3104 compounds were successfully docked and scored by the Surflex docking tool and the top scoring 2%, 62 compounds, were selected for procurement and testing. Of this number, 45 compounds have been obtained and tested, the remaining 17 compounds were no longer available from any supplier and have been slated for synthesis and testing in future studies. The compounds were tested at 500 μM concentration (250 μM if very poorly soluble) and a percentage inhibition was obtained. Compounds showing greater than 30% inhibition were taken into crystallography trials. Although this level of activity is slight when considering a standard high-throughput screen, it is an acceptable standard when dealing with fragment-like compounds, as was the case in these studies. As discussed in Chapter 1, with fragments it is more appropriate to consider binding efficiency rather than absolute binding affinity when determining which compounds to advance to further studies.

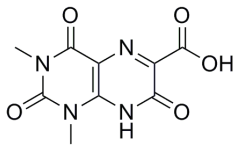
Twelve compounds met the activity requirement for further investigation and are shown in Figure 4.6. This corresponds to a hit rate of 26%, which is above average for a study of this nature. The addition of a pharmacophore filter prior to docking is most likely responsible for the increased hit rate over standard high-throughput virtual screening studies. All compounds shown in Figure 4.6 have been advanced to crystallography trials, the results of which are pending. It is noted that the hit compounds from the first round of virtual screening bear a close resemblance to the pterin substrate. Again, this can almost certainly be attributed to the pharmacophore filter employed prior to molecular docking. Figure 4.7 shows one of the pharmacophore hits docked into the pterin binding site by Surflex. The key pterin binding interactions are closely matched by the compound shown.

### 4.3.2 Fragment-Based Virtual Screen

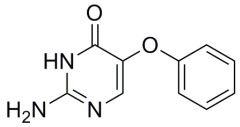
Unfortunately, the hit compounds from the first round of virtual screening, most likely due to their planar structure and aromatic stacking ability, are poorly soluble; making testing in enzyme assay and crystallography studies difficult. The low solubility of the hit compounds from round 1 was also felt to be a poor predictor for whole-cell biological activity and in vivo activity (studies not yet performed). Additionally, the first round of virtual screening did not yield the novel scaffolds for pterin site binding agents that we had hoped to find, due to the high degree of similarity between these hit compounds and the pterin substrate. For this reason, a second round of virtual



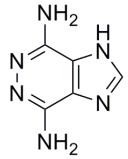
Lee-1014
62% inhibition, 500uM



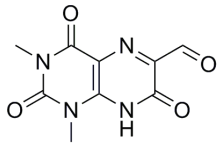
Lee-1022
32.4% inhibition, 500uM



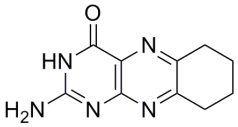
Lee-1023
37% inhibition, 500uM



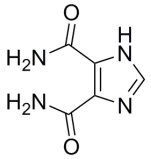
Lee-1027
32% inhibition, 500uM



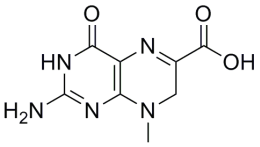
Lee-1029
44.1% inhibition, 500uM



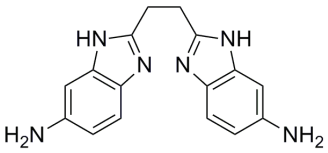
Lee-1032
31% inhibition, 500uM



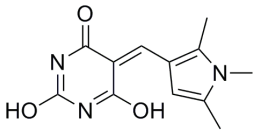
Lee-1035
32% inhibition, 500uM



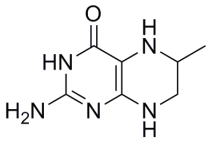
Lee-1111
92.8% inhibition, 250uM



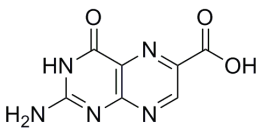
Lee-1176
61.2% inhibition, 500uM



Lee-1153
35.2% inhibition, 500uM



Lee-1331
67% inhibition, 500uM



Lee-1181
36.5% inhibition, 250uM

**Figure 4.6. Hits from Pharmacophore-Based Virtual Screening with Enzyme Activity Shown**

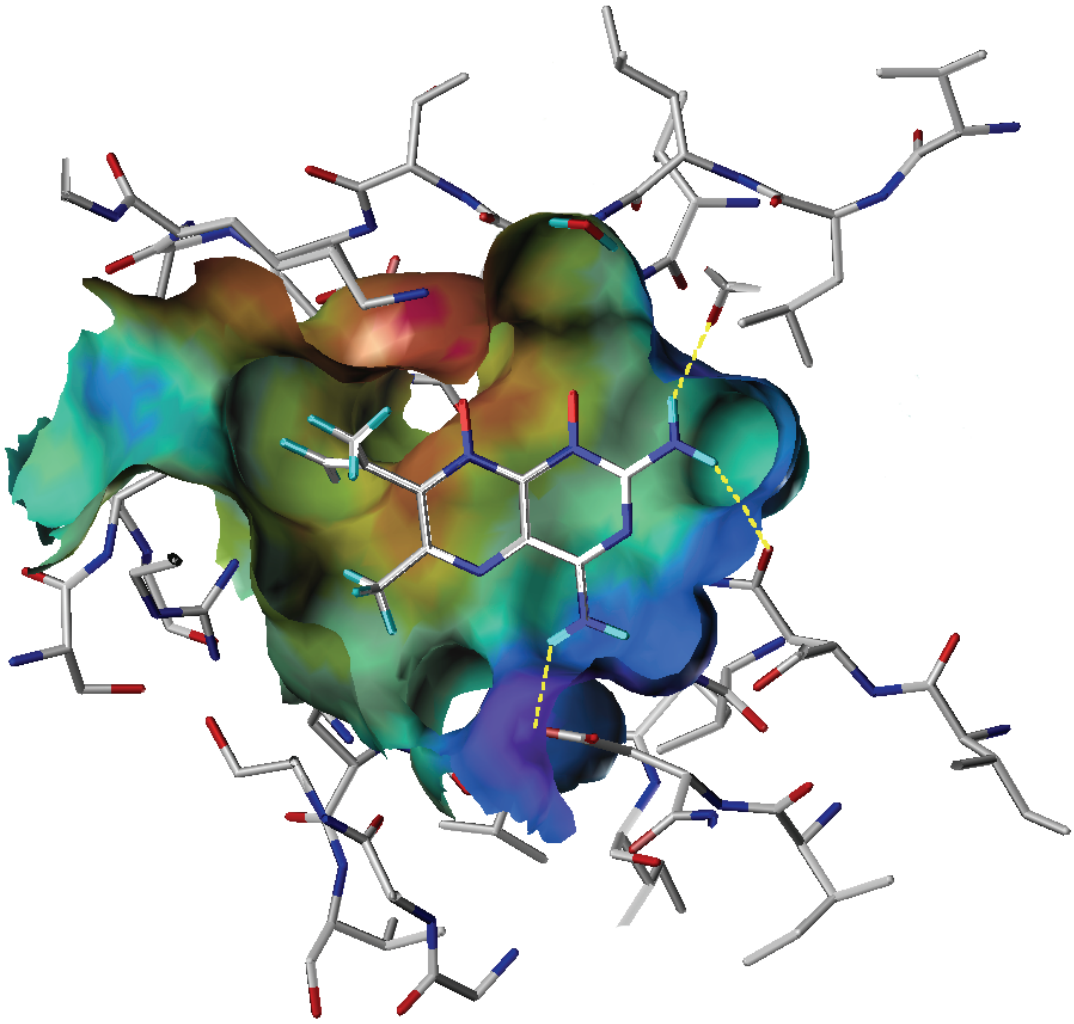


**Figure 4.7. Pharmacophore Hit Shown Docked into DHPS Pterin Site**

screening was performed, with tighter fragment constraints and no pharmacophore constraints.

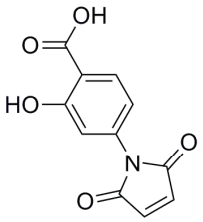
Due to computational limitations, the complete ZINC screening set could not feasibly be docked within a reasonable time frame. Therefore, for this round of virtual screening, a subset of specific vendors was selected and their screening libraries were obtained and docked (see methods above). The fragment constraints were tightened to a maximum molecular weight of 300 and no more than 5 rotatable bonds. In an attempt to improve the successful acquisition rate, the most updated screening libraries from each vendor used were obtained directly from the vendors and built as described in the methods section. The total number of compounds in the fragment sets docked was over 300,000. Using the Glide docking program, we successfully docked nearly 285,000 fragment compounds into the DHPS pterin site. A merged hit list of the top 1% of scored compounds from each supplier contained 2845 fragment compounds. Diversity filtering using the methods described above resulted in 54 clusters. The highest scoring compound was selected from each cluster for testing in the enzyme assay.

31 of the 54 fragment compounds from the second round of virtual screening have been successfully procured and tested to date. An additional 9 compounds are available from the suppliers, but at significant expense and therefore have not yet been ordered and tested. 8 compounds were rejected for ordering and testing due to close similarity to compounds already tested, and 1 fragment compound was a duplicate hit from the first round of virtual screening. 5 compounds were no longer available from any supplier and have been slated for in-house synthesis and testing. This corresponds to a 10% acquisition failure rate for the second round of virtual screening, compared with the 27% failure rate (15 or 62 compounds) for the first round. This significant improvement in successful acquisition of screening compounds is probably due to our use of the most current screening libraries from each vendor as well as only screening libraries from vendors with a proven track record from the first round of virtual screening.

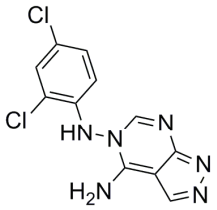
Of the 31 fragment compounds tested for activity in the second round of virtual screening, 3 compounds showed activity above the cut-off of 30% inhibition at 500μM and have been advanced into crystallography trials. This corresponds to a 10% hit rate, which is closer to the hit rates usually reported in high-throughput virtual screening studies. The compounds with measured activity are shown in Figure 4.8. These compounds bear much less resemblance to the pterin substrate and also have improved solubility over the hit compounds from the first round of virtual screening.

### 4.3.3 Comparison of Screening Methods and Results

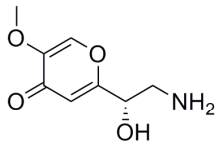
In this study we compared two separate methods for the virtual screening of a large number of compounds, on the order of several million, against a target. To compare the two methods we looked at several factors including: ease of use, computational expense and time requirements, and quality and character of the hit




Lee-1304
38.5% inhibition, 500uM




Lee-1324
46.4% inhibition, 500uM




Lee-1337
36.6% inhibition, 500uM

**Figure 4.8. Hits from Fragment-Based Virtual Screening with Enzyme Activity**

compounds. The obvious advantage of the virtual screening method that employed pharmacophore filtering (round 1) is that a much larger number of compounds, in this case nearly 5 million, could be screened within a reasonable amount of time using the computational resources available to our lab (4 processor batching capabilities on a Linux workstation running RHEL v4). A reasonable amount of time, in our case, was 2 to 4 weeks. This does not include the several weeks it took to create the Unity databases prior to the pharmacophore filtering step. Another advantage is that the hit rate was greater with the pharmacophore filtering method over the full fragment library docking method (26% versus 10%). The disadvantage, as mentioned above, to the pharmacophore filtering method is that, due to the nature of the pharmacophore filter, the hit compounds were all very similar to the pterin substrate. This similarity, unfortunately, included poor solubility. Another obvious ramification of this similarity is in terms of intellectual space and patentability.

In contrast to the pharmacophore filtering method, the fragment method, which involved only the application of a molecular weight and rotatable bond filter, yielded hit compounds that were significantly different that the pterin substrate with improved solubility. However, the elimination of the pharmacophore filtering step made this method more rigorous as we were required to explicitly dock and score all the compounds passing the fragment filter. Even by eliminating unreliable overseas vendors and employing a stricter fragment filter (300 Daltons versus 350 Daltons), it was still necessary to dock nearly 300,000 compounds, nearly 2 orders of magnitude greater than the number of compounds we explicitly docked in the pharmacophore filtering method. Obviously, this method requires a much longer period of time to complete, in our case nearly 8 weeks. As mentioned above, we also observed a significantly lower hit rate with this method. Finally, after successfully docking the fragments in round 2, the number of top scoring hits, even when selecting only 1% as opposed to 2% selected in round 1, was unmanageable in terms of our assay abilities as well as our acquisition budget, making it necessary to employ a further diversity filtering post-processing step.

### 4.3.4 Structure-Activity Relationship Studies

An analysis of the hit compounds from the first round of virtual screening has allowed us to develop a preliminary structure-activity relationship. The pharmacophore map shown in Figure 4.9 shows the observations that have been made based upon the activity of the pterin-like hits from the pharmacophore based search.

## 4.4 Summary

The results of two high-throughput fragment-based virtual screens against the bacterial target dihydropteroate synthase (DHPS) from *B. anthracis* have been reported. Molecular docking and scoring was performed in order to identify novel compounds and potential scaffolds targeting the pterin binding sub-site of DHPS. Pharmacophore filtering prior to docking was employed in the first round of virtual screening and compared to hit results from the second round which did not involve pharmacophore filtering. Although pre-filtering using pharmacophore constraints allowed the screening of compounds on the order of several million, the hit results were all limited to pterin-like compounds with limited solubility and little room for expansion into a novel scaffold area.

The second round of fragment-based virtual screening was much more computationally intensive; having forewent the pharmacophore filter and limited the number of compounds that were able to be screened to several hundred thousand. However, the hit compounds were all unique when compared to the pterin scaffold and displayed much greater solubility. Hit rates from the first round were much better than the second round (27% versus 10%), due to the pharmacophore match constraint applied in the first round. We also noted a lower successful compound acquisition rate for the ZINC compounds used in the first round of virtual screening when compared to building screening sets obtained directly from the supplier, as was done in the second round of virtual screening. Ultimately, 15 compounds met our activity cut-off from the enzyme assay employed in this study and were taken into crystallography trials.

A preliminary structure-activity analysis of the compounds from Round 1 of the virtual screening has been presented. Utilizing the activity information gained from the first round of virtual screening and the potentially novel scaffolds identified in the second round, we hope to develop a series of unique, DHPS pterin-site binding agents with potent activity against a broad range of gram-positive and gram-negative organisms.

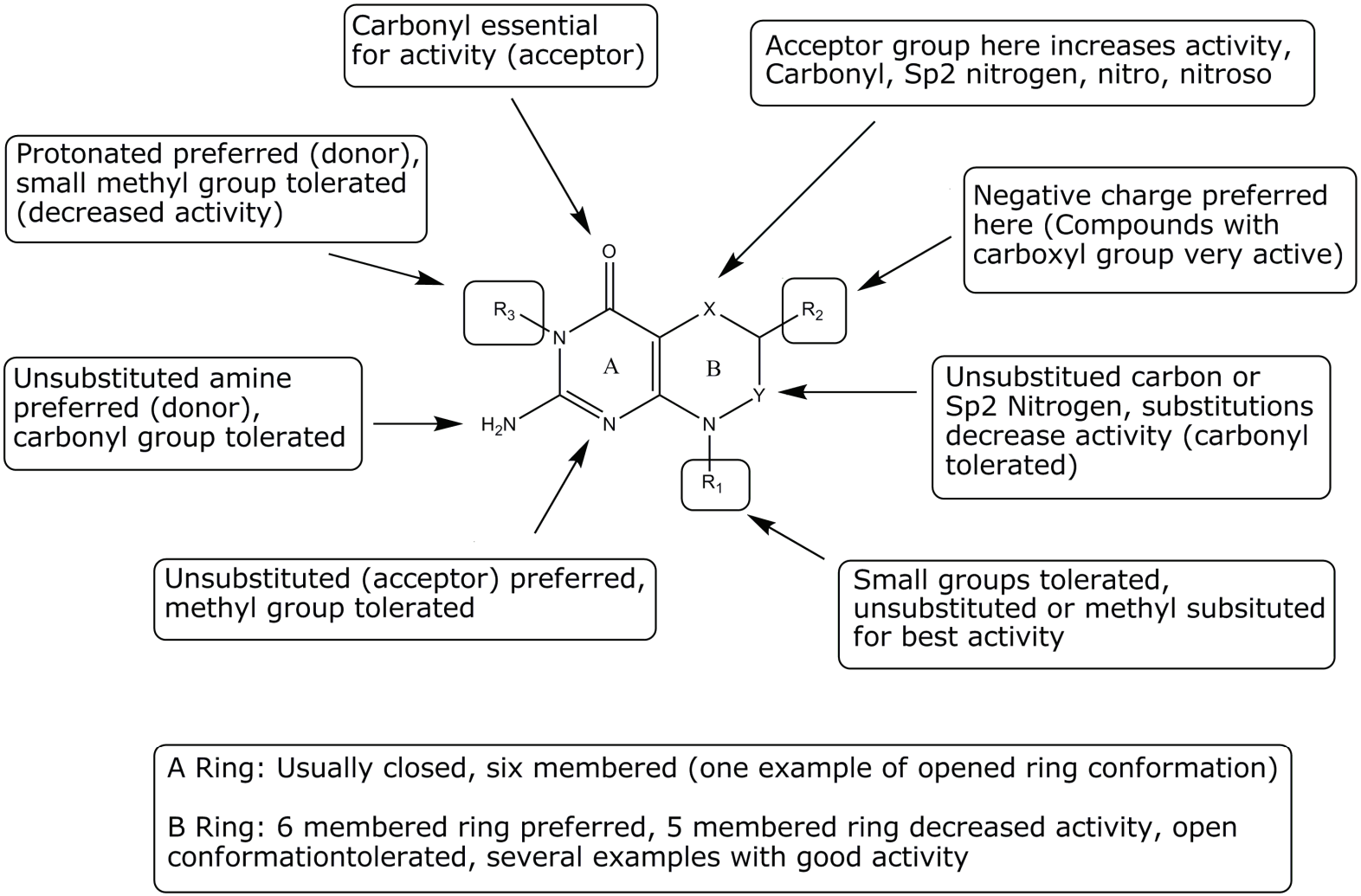



**Figure 4.9. Pharmacophore Map Based upon DHPS Screening Hit Activity**

# CHAPTER 5. LIGAND-BASED DESIGN OF NOVEL ANTITUBERCULAR AGENTS[1]

## 5.1 Introduction

### 5.1.1 The Tuberculosis Bacilli as a Target for Antimicrobial Drug Design

There is an urgent need today for new anti-tuberculosis agents with novel mechanisms of action. The global incidence of tuberculosis continues to rise, with a third of the world's population currently infected, yet there have been no new classes of antimycobacterial agents approved for use in forty years.[225] The efficacy of the currently available agents used in standard Tuberculosis (TB) treatment regimens is severely limited by several factors; including long treatment regimens, multiple drug treatment regimens, drug interactions, and drug resistance. The emergence of resistance, particularly Multi-Drug Resistant Tuberculosis (MDR-TB) and Extensively Drug Resistant Tuberculosis (XDR-TB), is particularly concerning. A recent report released by the World Health Organization estimated that the incidence of TB drug resistance (resistance to one drug in standard TB regimen) was as high as 57% in some countries, while multi-drug resistance was 14%.[226] Novel agents are needed that can bypass resistance mechanisms, that can treat the latent phase of infection shortening the duration of tuberculosis treatment, and that are compatible with HIV co-therapy by possessing low drug interaction rates.[227,228]

### 5.1.2 Nitrofuran Antituberculosis Agents

Toward these goals, our laboratory has been developing a series of nitrofuranyl compounds with potent whole-cell activity against *M. tuberculosis*.[229-235] Figure 5.1 shows the three major scaffolds in the nitrofuran series that have been examined so far. The series originated from a screen for TB cell wall inhibitors that produced a nitrofuran hit with a respectable MIC activity and low molecular weight.[229] Subsequent lead optimization efforts led to compounds with activity against the tuberculosis bacilli falling into the nanomolar range. Importantly, these compounds exhibit activity against both actively growing and latent bacilli, which is believed to be a beneficial attribute of potential new anti-tuberculosis agents.[234] Although the *in vitro* activity looks very promising for this nitrofuran series, poor solubility and metabolic instability have necessitated the development of further generations of nitrofuran agents that can overcome these issues. Ligand-based drug design techniques were employed to guide the synthesis of future generations of nitrofuran compounds, as described herein.

---

[1] Adapted by permission. Hevener, K. E.; Ball, D. M.; Buolamwini, J. K.; Lee, R. E. Quantitative structure-activity relationship studies on nitrofuranyl anti-tubercular agents. *Bioorg Med Chem* **2008,** 16, 8042-53.

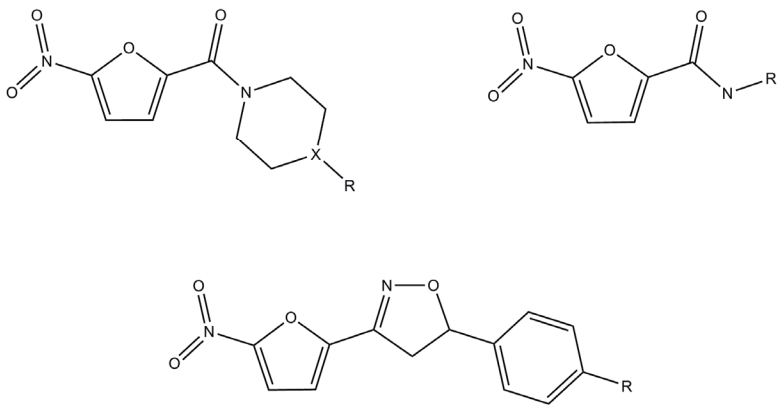


**Figure 5.1. Major Scaffolds of the Nitrofuran Compounds**

### 5.1.3 Nitrofuran QSAR Studies

Quantitative Structure-Activity Relationship (QSAR) techniques are methods used to correlate physicochemical descriptors from a set of related compounds to their known molecular activity or molecular property values.[30] QSAR models are a useful method of ligand-based drug design when the molecular target for the compounds being investigated is either unknown or has not been structurally resolved. The descriptors used to develop QSAR models can range from molecular descriptors for lipophilicity (cLogP and LogD)[31,32], steric bulk (Molar Refractivity, volume)[33], and electrostatics (polar surface area, Coulombic charges, dipole moments)[34] to 3-dimensional descriptors that involve alignment of the compounds, and calculating steric and electrostatic values using charged probe atoms at grid lattice points (CoMFA)[35] or 3-D similarity indices (CoMSIA).[36] Several quantitative structure-activity relationship models were developed in this study. Different molecular alignment rules were investigated in order to obtain models with high predictivity. Compounds with ionizable functional groups were investigated in their charged and uncharged states. Descriptors including cLogP, LogD, molar refractivity (CMR), polar surface area (PSA), and 3D CoMFA and CoMSIA variables were investigated for their ability to predict and correctly rank whole cell MIC activity using the method of Partial Least Squares, PLS.[41]

Since the activity data utilized in this 3D-QSAR study is whole cell activity expressed as the Minimum Inhibitory Concentration (MIC, see experimental section), it is assumed that the activity reflects several processes in addition to compound binding to the biomolecular target. Compound solubility, mycobacterial cell entry (i.e. passive diffusion or active transport), and stability to TB metabolism may all contribute to the whole cell activity. Additionally, these nitro-aromatic compounds are pro-drugs and must be metabolically activated by TB nitro reductase enzymes as already demonstrated for nitroimidazole agents PA824 and OPC67683 that are currently in clinical development.[236-238] The activated form is then believed to interact with its ultimate molecular target. Because of this multistep process, the development of reliable QSAR

models using whole cell activity is considered to be a difficult undertaking. However, several groups have reported success in the development of 3D-QSAR models using whole cell antimicrobial and antitubercular activity recently.[239-242] We have attempted to account for some of the processes mentioned above by investigating the addition of molecular descriptors that may be important factors for cell entry including lipophilicity and steric bulk to our 3D-QSAR models and testing the effects of ionized versus neutral compounds on the 3D-QSAR model's validity and predictive power.

## 5.2 QSAR Methods

### 5.2.1 Training and Test Set Preparation

Figure 5.2 graphically illustrates the general method followed for the development of the QSAR models in this study. We began with an initial set of 110 nitrofuran compounds with activity against *M. tuberculosis* (as determined by carefully standardized micro broth dilution MIC determination method, see experimental section). A test set of 15 compounds was selected from the remaining compounds for use in external validation. These test set compounds were selected such that their activity and physical properties (MW and cLogP) were broadly reflective of the training set characteristics (see experimental section). Tables 5.1 and 5.2 list the training set and test set nitrofuran compounds used in this study, respectively, along with their calculated molecular descriptors and biological activity. MIC activity originally determined in μg/mL were converted to micromolar values (μM/mL) and then converted to a pMIC value by taking Log (1/MIC). The pMIC values were used as the dependent variable in all PLS models subsequently developed. As a general rule, for a reliable 3D-QSAR model the spread of activity should cover at least 3 log units and there ideally should be a minimum of 15 to 20 compounds in the training set.[243] The activity range of the nitrofuran compounds ranged from 0.73 to 5.73 pMIC units (see Table 5.1), a full 5 log activity distribution, and there were 95 compounds in the training set. Figure 5.3 shows the training set and test set compounds distributed by their lipophilicity (cLogP) and molecular weight. The compounds are colored by activity. Importantly, it should be noted from this preliminary analysis that there is a correlation of increasing activity with molecular weight but no correlation with increasing cLogP, which may be expected for mycobacterial entry. We attribute the correlation with increasing molecular weight to the non-random nature of the data set, as these compounds result from the systematic medicinal chemistry development of the series from a low molecular weight, lower potency screening hit to high potency, higher molecular weight optimized compounds.

When designing a 3D-QSAR model using Comparative Molecular Field Analysis (CoMFA) or Comparative Molecular Similarity Indices Analysis (CoMSIA) the compounds in the training and test sets must share a common alignment, assumed to be the active conformation, and have the atomic charges loaded by a reliable method.[244] The compounds used in this study were built using the Sybyl Molecular Modeling

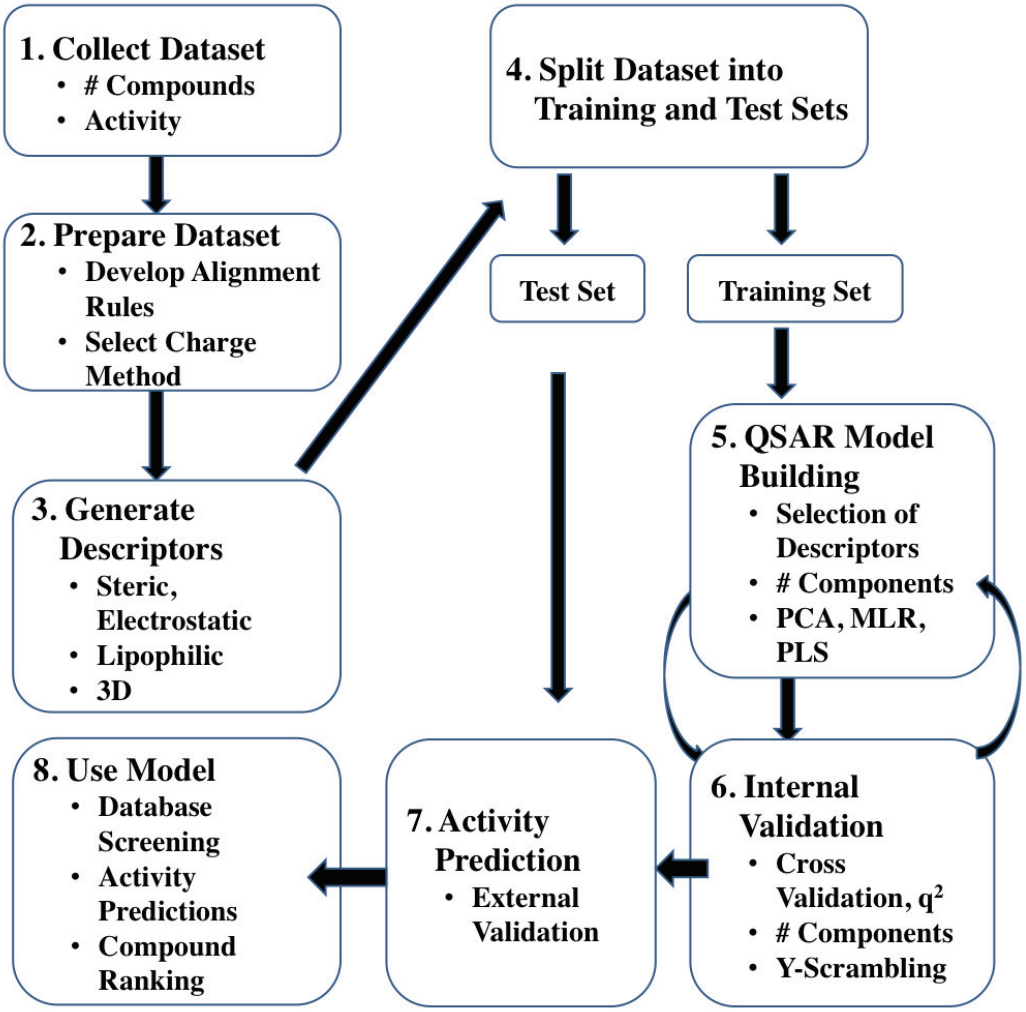



**Figure 5.2. QSAR Project Flowchart**

**Table 5.1. Physicochemical Properties and Activity of Training Set Compounds**

| Compound | Molecular Weight | PSA[a](A$^2$) | cLogP[b] | LogD$_{7.4}$[c] | CMR[b] | MIC(μg/mL) | pMIC[d] |
|---|---|---|---|---|---|---|---|
| $L_1$ | 290.314 | 137.057 | 1.84 | 1.50 | 7.488 | 3.1 | 1.9715 |
| $L_2$ | 232.192 | 155.828 | 1.68 | 1.95 | 5.893 | 0.8 | 2.4628 |
| $L_3$ | 276.220 | 145.403 | 2.57 | 2.77 | 6.903 | 0.4 | 2.8392 |
| $L_4$ | 517.454 | 172.693 | 5.84 | 6.19 | 12.725 | 0.025 | 4.3159 |
| $L_5$ | 382.342 | 163.142 | 4.12 | 4.20 | 10.031 | 0.003 | 5.1053 |
| $L_6$ | 400.306 | 130.996 | 3.39 | 3.33 | 8.852 | 0.0008 | 5.6993 |
| $L_7$ | 341.361 | 152.959 | 3.43 | 3.61 | 9.398 | 0.00156 | 5.3401 |
| $L_8$ | 252.266 | 146.082 | 1.89 | 1.70 | 6.451 | 3.1 | 1.9105 |
| $L_9$ | 236.181 | 174.535 | 0.36 | 0.50 | 5.571 | 6.25 | 1.5774 |
| $L_{10}$ | 257.202 | 218.788 | 1.71 | 1.76 | 6.371 | 0.8 | 2.5072 |
| $L_{11}$ | 276.245 | 165.292 | 1.62 | 1.31 | 6.974 | 0.1 | 3.4413 |
| $L_{12}$ | 280.664 | 153.319 | 2.30 | 2.08 | 6.849 | 1.6 | 2.2441 |
| $L_{13}$ | 306.271 | 173.676 | 1.49 | 1.05 | 7.591 | 0.4 | 2.8840 |
| $L_{14}$ | 306.271 | 164.455 | 1.49 | 1.05 | 7.591 | 0.2 | 3.1851 |
| $L_{15}$ | 336.297 | 164.287 | 1.37 | 0.80 | 8.208 | 0.8 | 2.6236 |
| $L_{16}$ | 286.283 | 142.587 | 2.54 | 2.41 | 7.571 | 3.1 | 1.9654 |
| $L_{17}$ | 317.297 | 170.828 | 1.56 | 1.87 | 8.093 | 3.13 | 2.0059 |
| $L_{18}$ | 330.339 | 157.250 | 1.72 | 1.43 | 8.772 | 12.5 | 1.4220 |
| $L_{19}$ | 406.434 | 155.828 | 3.45 | 3.00 | 11.284 | 0.8 | 2.7059 |
| $L_{20}$ | 405.446 | 155.828 | 4.63 | 4.88 | 11.379 | 3.13 | 2.1124 |
| $L_{21}$ | 272.256 | 144.663 | 2.12 | 2.01 | 7.107 | 3.1 | 1.9436 |
| $L_{22}$ | 330.339 | 157.180 | 1.72 | 1.44 | 8.772 | 12.5 | 1.4220 |
| $L_{23}$ | 406.434 | 155.828 | 3.45 | 3.01 | 11.284 | 12.5 | 1.5121 |
| $L_{24}$ | 405.446 | 155.828 | 4.63 | 4.88 | 11.379 | 12.5 | 1.5110 |
| $L_{25}$ | 393.396 | 164.787 | 3.17 | 3.51 | 10.609 | 6.25 | 1.7989 |
| $L_{26}$ | 234.168 | 203.256 | -0.28 | 0.02 | 5.471 | 6.25 | 1.5736 |
| $L_{27}$ | 314.336 | 110.259 | 2.82 | 2.70 | 8.500 | 0.4 | 2.8953 |
| $L_{28}$ | 276.245 | 165.039 | 1.62 | 1.31 | 6.974 | 1.6 | 2.2372 |
| $L_{29}$ | 306.271 | 164.057 | 1.84 | 1.02 | 7.590 | 1.2 | 2.4069 |
| $L_{30}$ | 292.244 | 212.380 | 1.23 | 1.02 | 7.127 | 0.39 | 2.8747 |
| $L_{31}$ | 288.255 | 171.250 | 1.17 | 0.94 | 7.261 | 9.38 | 1.4876 |
| $L_{32}$ | 290.228 | 195.200 | 1.53 | 1.24 | 6.950 | 0.15 | 3.2866 |
| $L_{33}$ | 332.308 | 136.172 | 2.06 | 1.59 | 8.341 | 0.1 | 3.5215 |
| $L_{34}$ | 324.309 | 221.944 | 0.45 | 0.34 | 7.694 | 0.17 | 3.2805 |
| $L_{35}$ | 331.323 | 168.233 | 1.63 | 1.48 | 8.557 | 0.4 | 2.9182 |
| $L_{36}$ | 344.365 | 154.693 | 1.79 | 1.00 | 9.236 | 0.4 | 2.9350 |
| $L_{37}$ | 420.461 | 153.312 | 3.52 | 2.56 | 11.747 | 0.0125 | 4.5268 |
| $L_{38}$ | 344.365 | 154.388 | 1.79 | 1.00 | 9.236 | 6.25 | 1.7411 |

**Table 5.1 (continued).**

| Compound | Molecular Weight | PSA[a](A$^2$) | cLogP[b] | LogD$_{7.4}$[c] | CMR[b] | MIC(µg/mL) | pMIC[d] |
|---|---|---|---|---|---|---|---|
| $L_{39}$ | 419.473 | 153.312 | 4.70 | 4.49 | 11.843 | 1.56 | 2.4296 |
| $L_{40}$ | 260.245 | 155.967 | 2.03 | 1.81 | 6.821 | 1.6 | 2.2113 |
| $L_{41}$ | 266.637 | 155.828 | 2.24 | 2.47 | 6.385 | 0.8 | 2.5228 |
| $L_{42}$ | 419.473 | 153.312 | 4.7 | 4.49 | 11.843 | 0.8 | 2.7196 |
| $L_{43}$ | 420.461 | 153.310 | 3.52 | 2.56 | 11.747 | 0.05 | 3.9248 |
| $L_{44}$ | 331.323 | 172.571 | 1.35 | 1.31 | 8.557 | 1.56 | 2.3271 |
| $L_{45}$ | 260.245 | 146.467 | 2.07 | 1.97 | 6.821 | 3.1 | 1.9240 |
| $L_{46}$ | 289.287 | 153.314 | 2.03 | 1.83 | 7.654 | 0.4 | 2.8593 |
| $L_{47}$ | 280.664 | 153.346 | 2.30 | 2.08 | 6.849 | 0.2 | 3.1472 |
| $L_{48}$ | 320.297 | 166.871 | 1.77 | 1.31 | 8.055 | 0.4 | 2.9035 |
| $L_{49}$ | 314.217 | 153.346 | 2.67 | 2.45 | 6.868 | 0.05 | 3.7983 |
| $L_{50}$ | 264.209 | 153.346 | 1.90 | 1.70 | 6.373 | 1.56 | 2.2288 |
| $L_{51}$ | 264.209 | 153.346 | 1.90 | 1.70 | 6.373 | 0.8 | 2.5189 |
| $L_{52}$ | 347.389 | 181.002 | 2.35 | 2.06 | 9.210 | 1.56 | 2.3477 |
| $L_{53}$ | 379.388 | 222.205 | 0.45 | 0.48 | 9.276 | 50 | 0.8801 |
| $L_{54}$ | 330.339 | 185.480 | 1.41 | -0.23 | 8.772 | 0.8 | 2.6159 |
| $L_{55}$ | 384.429 | 153.312 | 2.51 | 1.08 | 10.490 | 0.05 | 3.8858 |
| $L_{56}$ | 438.452 | 153.345 | 3.68 | 3.17 | 11.763 | 0.1 | 3.6419 |
| $L_{57}$ | 362.356 | 155.962 | 1.95 | 1.55 | 9.252 | 1.56 | 2.3660 |
| $L_{58}$ | 365.379 | 180.785 | 2.51 | 2.20 | 9.226 | 1.56 | 2.3696 |
| $L_{59}$ | 319.292 | 117.883 | 2.16 | 2.13 | 7.960 | 0.2 | 3.2032 |
| $L_{60}$ | 349.314 | 168.194 | 1.79 | 1.62 | 8.572 | 1.56 | 2.3501 |
| $L_{61}$ | 437.463 | 153.312 | 4.86 | 4.63 | 11.858 | 0.4 | 3.0389 |
| $L_{62}$ | 359.333 | 158.130 | 1.24 | 0.62 | 9.060 | 1.6 | 2.1907 |
| $L_{63}$ | 248.192 | 211.631 | 1.29 | 1.64 | 6.047 | 0.2 | 3.2545 |
| $L_{64}$ | 402.401 | 176.891 | 1.98 | 1.90 | 10.353 | 0.0062 | 4.8123 |
| $L_{65}$ | 415.443 | 183.540 | 1.79 | 1.71 | 11.032 | 0.2 | 3.3175 |
| $L_{66}$ | 415.443 | 175.337 | 1.62 | 1.67 | 11.032 | 0.8 | 2.7154 |
| $L_{67}$ | 293.255 | 239.058 | 2.56 | 1.92 | 6.870 | 50 | 0.7698 |
| $L_{68}$ | 267.236 | 196.771 | 2.94 | 2.38 | 6.293 | 50 | 0.7296 |
| $L_{69}$ | 325.382 | 106.482 | 3.62 | 4.25 | 9.216 | 25 | 1.1145 |
| $L_{70}$ | 446.498 | 119.369 | 3.73 | 2.81 | 12.498 | 0.0125 | 4.5529 |
| $L_{71}$ | 388.375 | 178.977 | 1.64 | 1.55 | 9.889 | 0.05 | 3.8903 |
| $L_{72}$ | 430.454 | 176.288 | 2.89 | 2.76 | 11.280 | 0.025 | 4.2360 |
| $L_{73}$ | 430.454 | 176.353 | 2.87 | 2.77 | 11.280 | 0.025 | 4.2360 |
| $L_{74}$ | 414.412 | 177.426 | 2.34 | 2.29 | 10.791 | 0.05 | 3.9185 |
| $L_{75}$ | 444.481 | 160.475 | 2.62 | 2.44 | 11.744 | 0.1 | 3.6479 |
| $L_{76}$ | 311.088 | 155.835 | 2.51 | 2.74 | 6.670 | 1.6 | 2.2888 |

**Table 5.1 (continued).**

| Compound | Molecular Weight | PSA[a]($A^2$) | cLogP[b] | $LogD_{7.4}$[c] | CMR[b] | MIC(μg/mL) | pMIC[d] |
|---|---|---|---|---|---|---|---|
| $L_{77}$ | 338.314 | 156.451 | 3.28 | 3.47 | 9.022 | 12.5 | 1.4324 |
| $L_{78}$ | 389.359 | 156.162 | 0.92 | 0.37 | 9.670 | 6.25 | 1.7945 |
| $L_{79}$ | 431.442 | 171.924 | 1.90 | 1.75 | 11.069 | 0.0008 | 5.7318 |
| $L_{80}$ | 357.380 | 141.100 | 2.41 | 2.57 | 9.283 | 6.25 | 1.7573 |
| $L_{81}$ | 434.488 | 153.312 | 3.63 | 1.66 | 12.211 | 0.8 | 2.7349 |
| $L_{82}$ | 416.428 | 170.810 | 2.09 | 1.95 | 10.816 | 0.1 | 3.6195 |
| $L_{83}$ | 429.470 | 171.300 | 1.73 | 1.71 | 11.496 | 0.4 | 3.0309 |
| $L_{84}$ | 421.449 | 162.679 | 2.90 | 2.23 | 11.536 | 0.0062 | 4.8324 |
| $L_{85}$ | 403.389 | 183.649 | 1.36 | 1.26 | 10.142 | 0.05 | 3.9068 |
| $L_{86}$ | 250.183 | 155.747 | 1.84 | 2.09 | 5.909 | 0.8 | 2.4952 |
| $L_{87}$ | 262.218 | 164.178 | 1.55 | 1.69 | 6.510 | 0.8 | 2.5156 |
| $L_{88}$ | 262.218 | 168.069 | 1.55 | 1.69 | 6.510 | 0.4 | 2.8166 |
| $L_{89}$ | 373.426 | 146.915 | 2.57 | 2.30 | 9.960 | 0.4 | 2.9701 |
| $L_{90}$ | 238.240 | 145.856 | 1.56 | 1.37 | 5.988 | 3.1 | 1.8857 |
| $L_{91}$ | 246.219 | 114.858 | 1.91 | 1.72 | 6.357 | 3.125 | 1.8965 |
| $L_{92}$ | 276.245 | 126.760 | 1.79 | 1.46 | 6.974 | 6.25 | 1.6454 |
| $L_{93}$ | 258.229 | 115.575 | 1.89 | 1.70 | 6.644 | 0.8 | 2.5089 |
| $L_{94}$ | 233.180 | 179.275 | 0.34 | 0.63 | 5.682 | 6.25 | 1.5718 |
| $L_{95}$ | 233.180 | 179.120 | 0.34 | 0.63 | 5.682 | 3.125 | 1.8728 |

a. Sybyl 8.0, Molecular Spreadsheet calculation, Tripos, Inc.[245]
b. ChemBioOffice Ultra 2008, CambridgeSoft, Inc.[246]
c. MarvinSketch, 4.1.13, ChemAxon, Inc.[247]
d. pMIC calculated as Log(1/MIC), where MIC values have been converted to μM/mL

**Table 5.2. Physicochemical Properties and Activity of Test Set Compounds**

| Test Set | Molecular Weight | PSA[a]($A^2$) | cLogP[b] | $LogD_{7.4}$[c] | CMR[b] | MIC(μg/mL) | pMIC[d] |
|---|---|---|---|---|---|---|---|
| $T_1$ | 334.325 | 153.027 | 4.09 | 4.31 | 9.399 | 0.025 | 4.1262 |
| $T_2$ | 393.396 | 169.032 | 2.50 | 3.51 | 10.610 | 0.4 | 2.9928 |
| $T_3$ | 247.207 | 169.364 | 0.83 | 0.39 | 6.146 | 0.8 | 2.4900 |
| $T_4$ | 319.293 | 218.116 | 2.75 | 2.43 | 7.796 | 1.6 | 2.3001 |
| $T_5$ | 264.194 | 200.702 | 1.05 | 0.96 | 6.088 | 1.6 | 2.2178 |
| $T_6$ | 290.271 | 168.402 | 1.90 | 1.56 | 7.438 | 0.8 | 2.5597 |
| $T_7$ | 260.245 | 140.827 | 2.07 | 1.97 | 6.821 | 1.6 | 2.2113 |
| $T_8$ | 330.339 | 228.151 | 0.98 | -1.41 | 8.772 | 0.8 | 2.6172 |
| $T_9$ | 347.298 | 146.305 | 1.33 | 1.01 | 8.455 | 1.56 | 2.3476 |
| $T_{10}$ | 279.272 | 208.089 | 1.88 | 1.99 | 6.894 | 50 | 0.7486 |
| $T_{11}$ | 370.402 | 120.369 | 2.00 | 1.24 | 9.986 | 0.05 | 3.8697 |
| $T_{12}$ | 341.381 | 122.884 | 2.36 | 2.28 | 9.249 | 6.25 | 1.7374 |
| $T_{13}$ | 233.180 | 179.557 | 1.06 | 1.33 | 5.682 | 3.125 | 1.8728 |
| $T_{14}$ | 222.158 | 231.397 | 0.49 | 0.97 | 5.112 | 6.25 | 1.5508 |
| $T_{15}$ | 214.132 | 204.941 | 0.31 | -0.47 | 4.499 | 0.4 | 2.7286 |

a. Sybyl 8.0, Molecular Spreadsheet calculation, Tripos, Inc.[245]
b. ChemBioOffice Ultra 2008, CambridgeSoft, Inc.[246]
c. MarvinSketch, 4.1.13, ChemAxon, Inc.[247]
d. pMIC calculated as Log(1/MIC), where MIC values have been converted to M/mL.

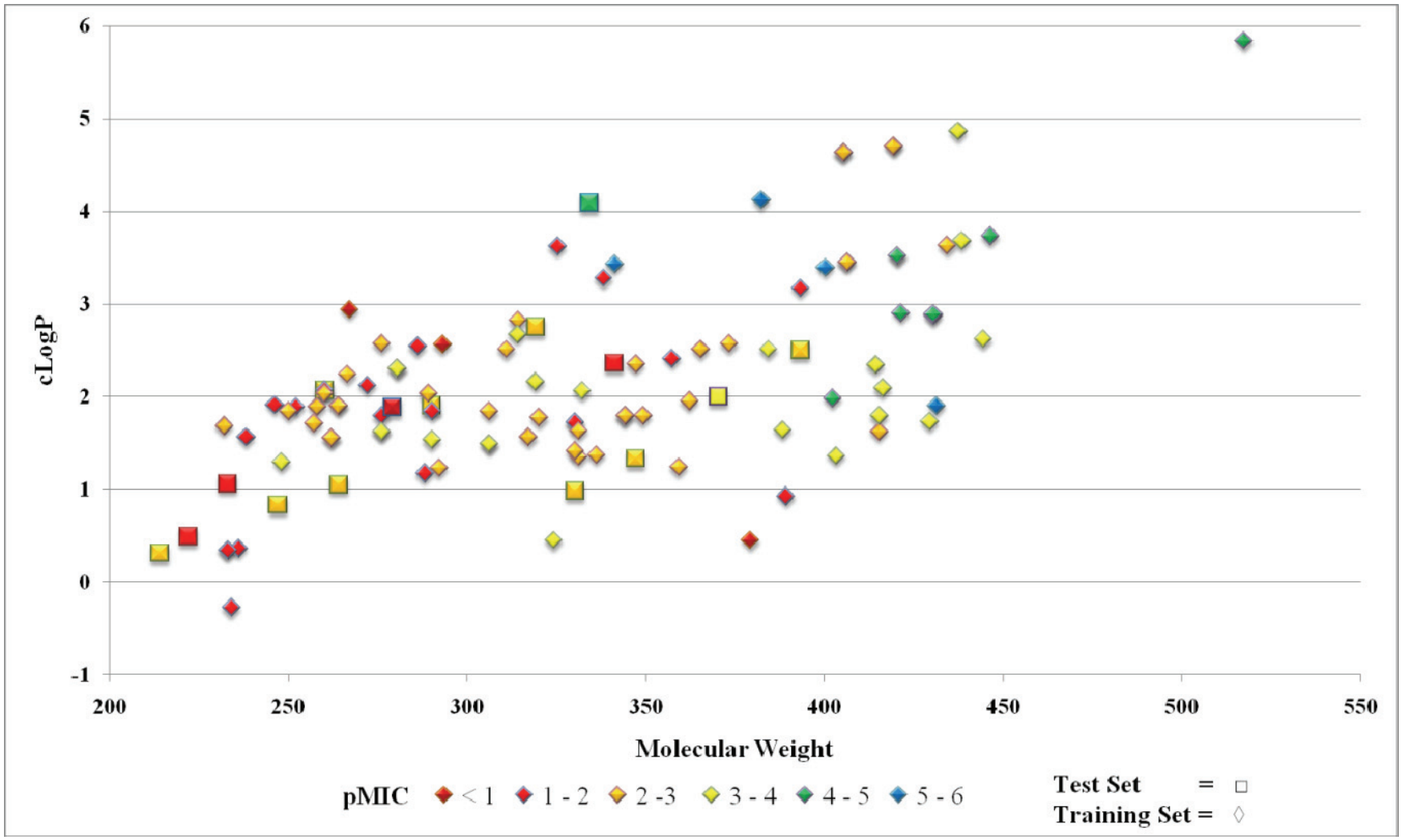



**Figure 5.3. Nitrofuran Training and Test Set Compounds by Physical Property and Activity**

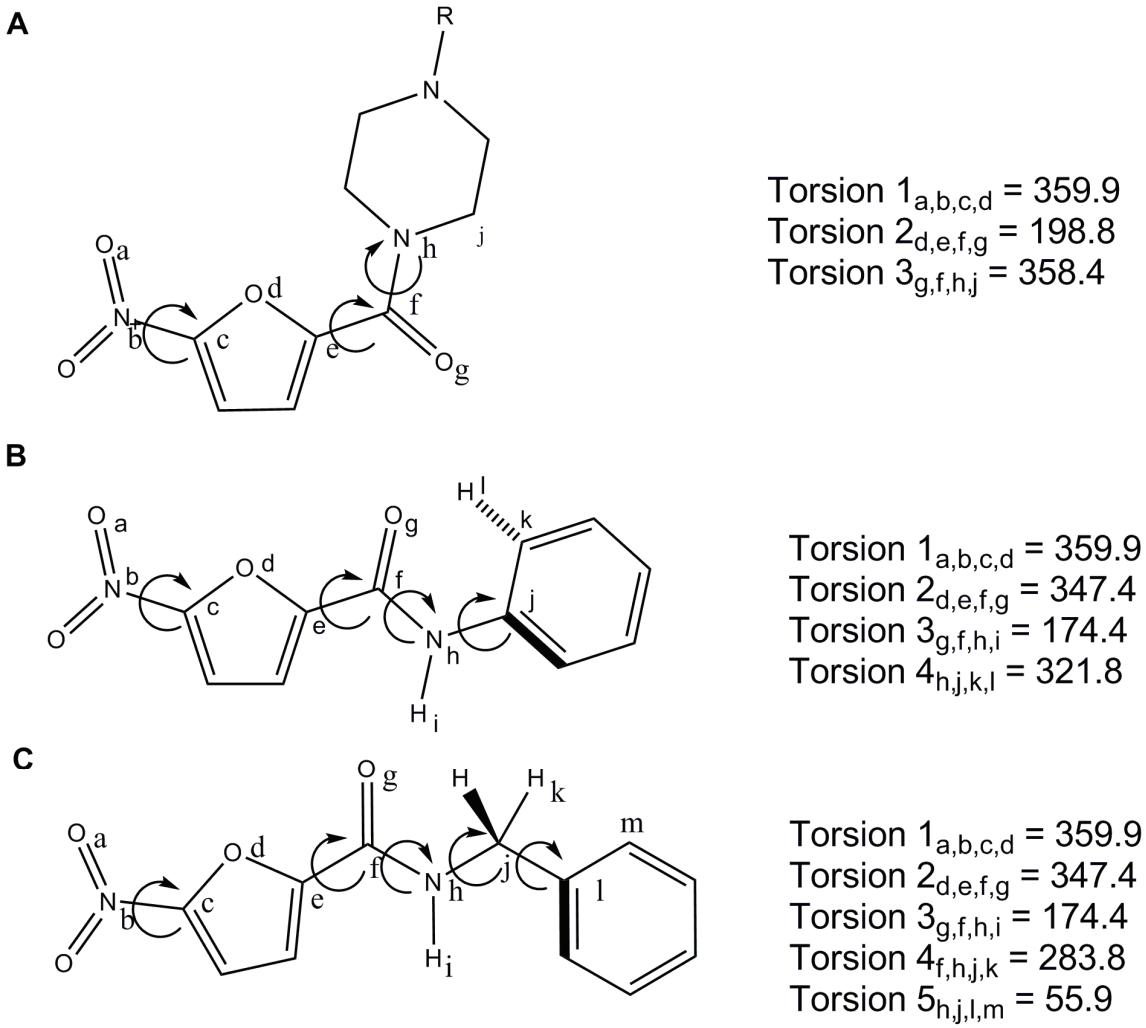
Package of Tripos, Inc.[245] The charges were loaded on all compounds in the training andtest sets using the PM3 semi-empirical method contained in the MOPAC suite.[248] Several of the nitrofuran compounds contained ionizable functional groups that would be expected to carry a charge at physiological pH. In order to account for this and to investigate the effect of protonating or de-protonating these functional groups on model predictivity, two sets of models were built for each alignment rule utilized. The first set of PLS models used all nitrofuran compounds in their neutral state and the cLogP molecular descriptor for lipophilicity (when a lipophilicity descriptor was used), the second set of PLS models used ionized nitrofuran compounds, as determined by a major microspecies calculation (discussed in the experimental section), and LogD as the lipophilic descriptor. Because the molecular target of the nitrofuran compounds is unknown and the active conformation remains unclear, multiple alignments for these compounds were studied in an attempt to generate the optimal PLS model in terms of activity prediction.

Alignment rules were determined by calculating energy minima using the Grid Search function of Sybyl and 10 degree increments against the rotatable bonds in our most active nitrofuran compounds from each representative scaffold. The first alignment method specified all nitrofuran compounds be aligned in the same orientation: a sterically unhindered trans-amide conformation shown in Figure 5.4, A. The second alignment method specified that the compounds were aligned to the minimum energy conformations of several of the more active nitrofuran compounds. Due to differences in the side chains and steric hindrance factors, the second method actually consisted of separate alignment rules for phenyl substituted, benzyl substituted, and hindered tertiary amide nitrofurans. Figure 5.4, B and C show the alignment rules adopted for unhindered phenyl and benzyl substituted nitrofurans. Sterically hindered tertiary amide nitrofurans were aligned using the rules shown in Figure 5.4, A, which conform more closely to the minimum energy conformation seen with these compounds and is the same rule adopted for all compounds in the first alignment method. We note that the selected conformation of our nitrofuran compounds in 5.4, B and C very closely aligns with the structure of PA824 determined in a recently solved crystal structure.[249]

Global molecular and 3D physicochemical descriptors were calculated for all compounds in the training and test set and used to develop the QSAR models (see experimental section). Lipophilicity descriptors included cLogP, LogD, and Polar Surface Area (PSA). Molecular volume and steric bulk were also investigated using molar refractivity (CMR) as a molecular descriptor. 3D-QSAR methods utilized were CoMFA and CoMSIA. The performance of the 3D models before and after the addition of various combinations of molecular descriptors was investigated.

### 5.2.2 QSAR Model Development

The QSAR models investigated in this study were built using the Molecular Spreadsheet tool in the Sybyl 8.0 suite of Tripos, Inc.[245] 3-dimensional descriptors were




Torsion $1_{a,b,c,d}$ = 359.9
Torsion $2_{d,e,f,g}$ = 198.8
Torsion $3_{g,f,h,j}$ = 358.4

Torsion $1_{a,b,c,d}$ = 359.9
Torsion $2_{d,e,f,g}$ = 347.4
Torsion $3_{g,f,h,i}$ = 174.4
Torsion $4_{h,j,k,l}$ = 321.8

Torsion $1_{a,b,c,d}$ = 359.9
Torsion $2_{d,e,f,g}$ = 347.4
Torsion $3_{g,f,h,i}$ = 174.4
Torsion $4_{f,h,j,k}$ = 283.8
Torsion $5_{h,j,l,m}$ = 55.9

**Figure 5.4. Nitrofuran Alignment Rules Used for QSAR Studies**

generated using both CoMFA and CoMSIA methods as described in the experimental section below. The effect of outlier removal, number of components, and incorporation of molecular descriptors in the 3D models were investigated for the CoMFA and CoMSIA models generated. The program SAMPLS was used to gauge the optimum number of components for each model during model development.[250] In order to avoid over- fitting the models, a higher component was only accepted and used if it resulted in an increase of greater than 10% to the cross-validated $r^2$ ($q^2$) values. Progressive scrambling was performed to confirm the optimum number of PLS components and dependent variable scrambling was done to check for chance correlation within the models generated.[251-253]

The best model was obtained using the following methodology: First, models were generated for each alignment and ionization rule using both CoMFA and CoMSIA fields without the addition of molecular descriptors or the removal of any outlier compounds. Next, the molecular descriptors cLogP, LogD, CMR, and PSA were investigated for their ability to improve the best CoMFA and CoMSIA models. Following this, the best performing CoMFA and CoMSIA models at this stage was optimized by the successive removal of outlier compounds (see discussion below) and finally by region focusing.[254]

### 5.2.3 QSAR Model Validation

The strength of all the models developed was evaluated by a number of validation methods, including internal cross-validation, and external test set predictions. The cross validation methods of Leave-One-Out (LOO) and Leave-Group-Out (10 compound groups) were chosen to generate cross validated $r^2$ ($q^2$) values and Standard Errors of Prediction (SEP). Bootstrapping (10 runs) was utilized to calculate confidence intervals for the $r^2$ and Standard Errors of Estimate (SEE). The equations for $q^2$ and standard errors are given below. Models generated were used to predict activity for the test set compounds and generate activity correlated $r^2$ values. Coefficient of determination, $r^2$, values and standard errors were generated for the final models developed. Models were considered questionable if the difference between cross-validated $r^2$ ($q^2$) and non-validated $r^2$ was > 0.3.[255]

$$q^2 = 1 - \frac{\sum_y (Y_{pred} - Y_{actual})^2}{\sum_y (Y_{pred} - Y_{mean})^2} \qquad \textbf{Equation 5.1}$$

where: $Y_{pred}$ = predicted activity, $Y_{actual}$ = experimental activity, and $Y_{mean}$ = the best estimate of the mean

$$SEE, SEP = \sqrt{\frac{PRESS}{n - c - 1}} \qquad \textbf{Equation 5.2}$$

where: n = number of compounds, c = number of components, and:

$$\mathrm{PRESS} = \sum_{y} (Y_{pred} - Y_{actual})^2$$ **Equation 5.3**

### 5.2.4 Experimental Methods

*Training and Test Set Preparation.* All nitrofuranyl compounds investigated in this QSAR study were originally synthesized and tested for activity in our lab.[229-231] Compounds were built using the Sybyl 8.0 molecular modeling package and charges were loaded using the PM3 semi-empirical method available in the MOPAC suite. The compounds were minimized using the Powell method with an initial Simplex optimization and gradient termination of 0.01 kcal/mol (500 maximum iterations). The global molecular descriptors cLogP and CMR were calculated using ChemBioOffice 2008.[246] Polar surface area was calculated using the molecular spreadsheet application in Sybyl 8.0.[245] LogD was calculated for compounds at pH 7.4 using the calculator plugin tool in Marvin 4.1.13.[247] Ionized compounds were identified by performing a major microspecies calculation on all compounds in the training and test set at pH 7.4 using the calculator plugin tool of Marvin 4.1.13.[247] All compounds were aligned manually as discussed above. The 15 test set compounds were chosen from the 110 nitrofuran compounds by selecting for diversity using the program, Selector.[256] Selector is an application available in the Sybyl 8.0 molecular modeling suite.[245] Atom pairs and 2D fingerprints were used to form 15 diversity clusters by hierarchical clustering. 1 compound was selected from each cluster, chosen to maximize the spread of activity data.

*QSAR Model Validation.* SAMPLS was used to initially select the optimum number of components used in the PLS models generated[250]; with the exception noted above that a higher component was selected only if it resulted in an increase in $q^2$ values of at least 10%. Group cross-validation used 10 groups in all cases. Bootstrapped results were obtained using 10 bootstrapping runs. The progressive scrambling stability test was performed up to 10 components using 50 scramblings, 10 maximum bins, and 2 minimum bins. The critical point was 0.85 and the seed was 12080.

*QSAR Model Development*. 3-D CoMFA descriptors were generated using c.3 probe atom with a +1 charge and a grid spaced at 2 Å and extending 4 Å beyond the compounds in all directions. Tripos Standard CoMFA steric and electrostatic fields were generated using a distance dependent dielectric, no smoothing, and cutoffs of 30 kcal/mol for each. CoMSIA similarity fields were calculated for steric, electrostatic, hydrophobic, h-bond donor, and h-bond acceptor using the default attenuation factor of 0.3. Partial Least Squares analysis was used to build models correlating descriptors to the dependent variable, pMIC. Optimum number of components was determined by SAMPLS, cross validation methods, and progressive scrambling. A column filtering value of 0.5 and CoMFA standard scaling was used in all PLS analyses. Region

focusing was performed by applying a discriminant power weighting factor of 0.3 and new grid spacing equal to the original.

*Antituberculosis Activity Testing.* MIC values were determined using the microbroth dilution method and were read by visual inspection. Two-fold serial dilutions of test compound were prepared in 96-well round bottom microtiter plates (Nunc, USA) in 100 µL of the 7H9 broth media (Difco Laboratories, MI, USA) supplemented with 10% Albumin-Dextrose Complex and 0.05% (v/v) Tween80. An equivalent volume (100 µL) of broth inocula containing approximately $10^5$ CFU/mL of *M. tuberculosis* H37Rv was added to each well to give final concentrations of test compound starting at 200 µg/mL. The plates were incubated aerobically at 37°C for 7 days and the MIC was recorded as the lowest concentration of drug which inhibited 90% of growth with respect to the no-drug control.

## 5.3. Results and Discussion

### 5.3.1 General Validation and Predictivity Results

Descriptions of the 3D-QSAR models built are given in Table 5.3; the validation data and predictive ability are shown in Table 5.4. PLS models which used CoMFA generated 3D descriptors generally outperformed models using CoMSIA 3D descriptors. It should be noted that all 5 CoMSIA fields were used in the PLS (steric, electrostatic, hydrophobic, h-bond donor, and h-bond acceptor) built in this study. The rules of alignment and ionization had a strong influence on the final performance of the models generated. Models using ionized nitrofuran compounds, Figure 5.5, generally performed worse than the neutral compound models, with the exception of model 2 and 10, both of which had higher test set $r^2$ and non-validated $r^2$ values, but lower internal validation, $q^2$, values. This may be reflective of the need for neutral compounds to passively diffuse into the mycobacterial cell, or possibly the binding of the nitrofuran compounds to their biomolecular target in a neutral state. Models generated using alignment 1, in which all nitrofuran compounds adopted the sterically unhindered trans-amide conformation, also performed significantly worse than those built using alignment 2, in which compounds adopted one of three minimum energy conformations. Test set activity predictions were particularly poor for the alignment 1 QSAR models, and the cross-validation also demonstrated that these were much weaker models compared with alignment 2 models. In light of this data, the decision was made to advance model 3 (CoMFA, alignment 2, neutral compounds) and model 7 (CoMSIA, alignment 2, neutral compounds) into the next stage of model development, which involved the investigation of molecular descriptors ability to improve the model's predictivity.

**Table 5.3. QSAR Model Descriptions**

| Model | Description | Alignment | Ionization | # Components | Outliers |
|---|---|---|---|---|---|
| 1 | CoMFA | 1 | No | 1 | 0 |
| 2 | CoMFA | 1 | Yes | 2 | 0 |
| 3 | CoMFA | 2 | No | 3 | 0 |
| 4 | CoMFA | 2 | Yes | 3 | 0 |
| 5 | CoMSIA | 1 | No | 2 | 0 |
| 6 | CoMSIA | 1 | Yes | 2 | 0 |
| 7 | CoMSIA | 2 | No | 3 | 0 |
| 8 | CoMSIA | 2 | Yes | 2 | 0 |
| 9 | CoMFA, cLogP | 2 | No | 3 | 0 |
| 10 | CoMFA, LogD | 2 | Yes | 4 | 0 |
| 11 | CoMFA, PSA | 2 | No | 3 | 0 |
| 12 | CoMFA, CMR | 2 | No | 3 | 0 |
| 13 | CoMFA, cLogP, CMR | 2 | No | 3 | 0 |
| 14 | CoMFA, cLogP, PSA | 2 | No | 3 | 0 |
| 15 | CoMFA, PSA, CMR | 2 | No | 4 | 0 |
| 16 | CoMFA, cLogP, PSA, CMR | 2 | No | 4 | 0 |
| 17 | CoMSIA, cLogP | 2 | No | 3 | 0 |
| 18 | CoMFA | 2 | No | 3 | 3 |
| 19 | CoMFA | 2 | No | 5 | 6 |
| 20 | CoMFA | 2 | No | 5 | 7 |
| 21 | CoMFA | 2 | No | 5 | 8 |
| 22 | CoMSIA | 2 | No | 3 | 6 |
| 23 | CoMFA (19) Region Focused | 2 | No | 5 | 6 |

**Table 5.4. QSAR Model Validation and Predictivity**

| Model | LOO Cross $q^2$ / SEP | Group Cross $q^2$ / SEP | Bootstrapped $r^2$ | Bootstrapped SEE | Non-Validated $r^2$ / SEE | Test Set $r^2$ / SEE |
|---|---|---|---|---|---|---|
| 1 | .166 / 1.009 | .162 / 1.012 | .414 ± .079 | .886 ± .393 | .294 / .928 | .118 / .831 |
| 2 | .139 / 1.030 | .130 / 1.036 | .471 ± .047 | .799 ± .326 | .355 / .842 | .293 / .750 |
| 3 | .286 / .935 | .279 / .930 | .741 ± .041 | .564 ± .279 | .650 / .655 | .769 / .456 |
| 4 | .235 / .968 | .236 / .974 | .683 ± .050 | .642 ± .340 | .580 / .717 | .648 / .591 |
| 5 | .167 / 1.014 | .187 / 1.001 | .523 ± .044 | .728 ± .298 | .425 / .842 | .567 / .611 |
| 6 | .153 / 1.022 | .127 / 1.038 | .557 ± .044 | .718 ± .301 | .451 / .823 | .417 / .613 |
| 7 | .240 / .964 | .215 / 1.004 | .690 ± .030 | .637 ± .313 | .588 / .710 | .786 / .497 |
| 8 | .205 / .981 | .203 / .982 | .563 ± .071 | .679 ± .285 | .451 / .816 | .441 / .667 |
| 9 | .326 / .908 | .320 / .913 | .683 ± .059 | .636 ± .227 | .558 / .735 | .528 / .746 |
| 10 | .264 / .954 | .238 / .971 | .705 ± .065 | .594 ± .293 | .588 / .714 | .697 / .556 |
| 11 | .265 / .949 | .261 / .951 | .645 ± .043 | .640 ± .232 | .559 / .735 | .609 / .601 |
| 12 | .311 / .918 | .314 / .916 | .690 ± .034 | .581 ± .204 | .633 / .670 | .737 / .498 |
| 13 | .304 / .923 | .295 / .929 | .632 ± .030 | .674 ± .202 | .552 / .740 | .514 / .781 |
| 14 | .296 / .928 | .305 / .922 | .594 ± .048 | .705 ± .242 | .486 / .793 | .525 / .757 |
| 15 | .284 / .941 | .288 / .938 | .742 ± .034 | .549 ± .240 | .636 / .671 | .717 / .533 |
| 16 | .308 / .925 | .326 / .913 | .680 ± .045 | .622 ± .242 | .578 / .723 | .419 / .836 |
| 17 | .402 / .855 | .358 / .887 | .627 ± .051 | .674 ± .278 | .601 / .698 | .559 / .618 |
| 18 | .448 / .732 | .420 / .750 | .794 ± .023 | .447 ± .175 | .725 / .516 | .756 / .474 |
| 19 | .530 / .664 | .537 / .660 | .923 ± .016 | .251 ± .097 | .856 / .368 | .781 / .561 |
| 20 | .559 / .643 | .588 / .621 | .919 ± .015 | .265 ± .116 | .884 / .330 | .734 / .612 |
| 21 | .581 / .631 | .600 / .616 | .926 ± .020 | .250 ± .116 | .896 / .315 | .754 / .584 |
| 22 | .573 / .668 | .589 / .607 | .768 ± .041 | .465 ± .149 | .708 / .512 | .781 / .493 |
| 23 | .585 / .625 | .587 / .623 | .903 ± .024 | .305 ± .127 | .845 / .381 | .769 / .524 |

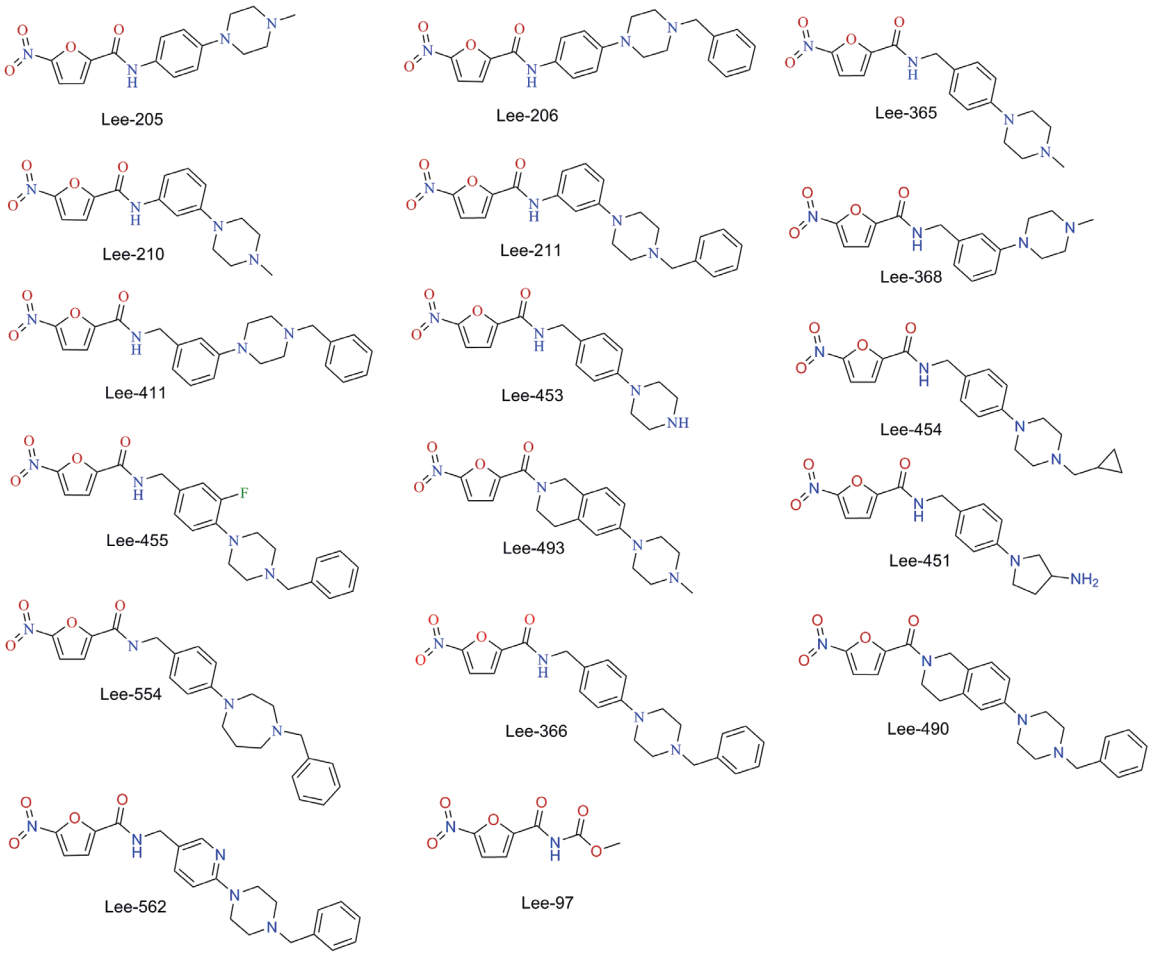



**Figure 5.5. Nitrofuran Compounds with Predicted Charge at Physiological pH[a,b]**

a. As determined by major microspecies calculation using MarvinSketch, v. 4.1.13. [247]
b. Physiological pH, 7.4

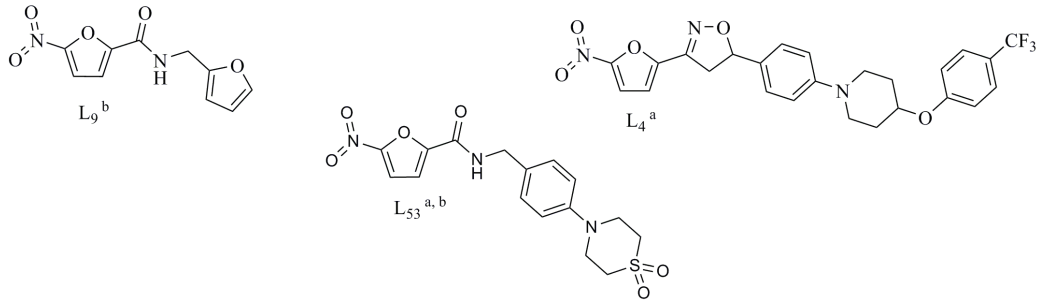
### 5.3.2 The Effects of Adding Global Molecular Descriptors

Global molecular descriptors were added to the 3D-QSAR models developed in an attempt to account for factors contributing to the MIC including, solubility and cell entry. The addition of cLogP to Model 3 led to a significant improvement in the cross-validated $r^2$ (internal validation), but a lower non-validated and bootstrapped $r^2$ (model 9). A similar result was seen when cLogP was added to CoMSIA fields in a reflective PLS analysis (model 17); the cross validated $r^2$ values were significantly higher, but the non-validated and test set $r^2$ values were not an improvement over model 7. The addition of LogD values to model 4 (in order to investigate ionization) had negligible effect on the internal validity or test set prediction of that model. Polar Surface Area (PSA) values added to model 3 had a negligible effect on internal validity of the model and worsened the predictivity, as seen by the decreased performance against the test set. The addition of CMR as a measure of steric bulk of the nitrofuran compounds led to slight improvements in the cross-validated $r^2$ values, but again, lower bootstrapped and testset $r^2$ values. Similarly, various combinations of the molecular descriptors, as shown in models 13 through 16, did not improve model 3 to any significant extent Ultimately, the models selected to proceed to step 3 (outlier investigation) were models 3 and 7, which do not incorporate any global molecular descriptors.

### 5.3.3 Outlier Compounds

Figure 5.6 shows outlier nitrofuran compounds, the removal of which improved the CoMFA and CoMSIA models discussed herein. Outlier compounds removed from each model were determined by analysis of a QQ plot generated by the QSAR analysis tool of Tripos, Inc. The QQ plot is essentially a normal probability plot of residuals, which is a validated method specifically developed to detect outliers.[255,257] Compounds with residuals that did not follow normal distribution were removed sequentially from the models developed, starting with the highest deviation from normal distribution. Model 18 was generated by removal of compounds $L_6$, $L_{64}$, and $L_{79}$, all with under-predicted activity. Model 19 was generated by removing 3 more compounds; $L_4$, $L_{53}$, and $L_{49}$. Subsequent outlier removal (model 20 and model 21) did not result in the improvement of the CoMFA models to a significant extent. It can be seen from the data given in Table 5.4 that the removal of 6 outliers was optimal in terms of predictive ability of the CoMFA models as demonstrated by the test set $r^2$ values. Although there was modest improvement in the internal validity (seen by cross-validated $r^2$ values for CoMFA) by removal of additional outlier compounds, there was negligible improvement to bootstrapping and non-validated $r^2$ values. CoMSIA model 22 was generated by removal of six compounds from CoMSIA model 7 again based upon the residual distribution. The CoMSIA outlier compounds are shown in Figure 5.6. Four of the six outlier compounds removed to generate CoMSIA model 22 were also outlier compounds from the CoMFA models (model 18 and model 19). CoMSIA model 22 showed significant improvement to both cross-validated and non-validated $r^2$ values but had little effect on the test set $r^2$ values, indicating an improvement in validity without affecting

**A. Compounds with over-predicted activity**




**B. Compounds with under-predicted activity**

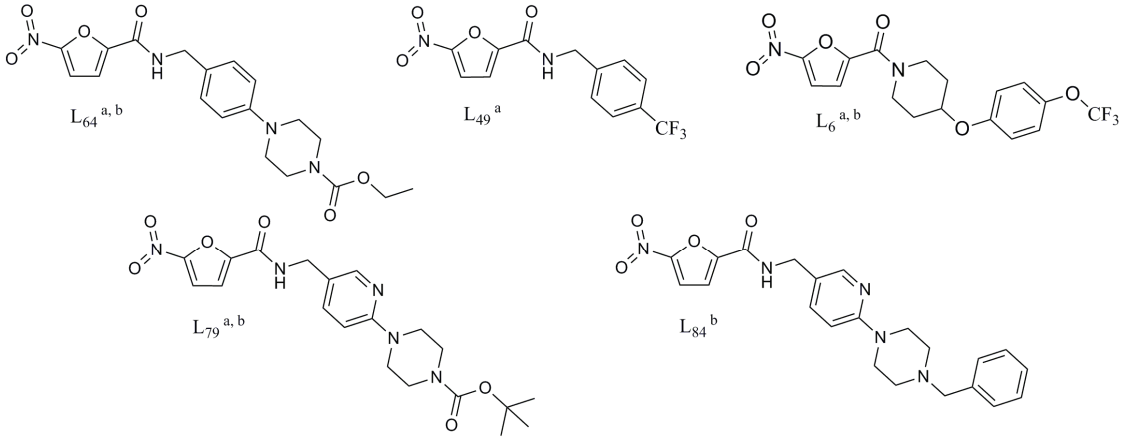



**Figure 5.6. Structures of Outlier Compounds**

a. Outliers from CoMFA Model 19.
b. Outliers from CoMSIA Model 22.

external predictivity. This model had comparable internal validity and test set predictivity to our best CoMFA model (model 19), but the bootstrapped and non-validated $r^2$ values were significantly lower. For this reason, model 19 (CoMFA, 6 outliers) was chosen to take to the final step in the 3D-QSAR development, region focusing.

The compounds in Figure 5.6 are sorted by whether their activity was over-predicted or under-predicted. Failure of these compounds to perform well in the QSAR models can be due to several factors, including inability to align correctly with the training set, inaccurate activity values, other processes not accounted for (i.e. active transport, prodrug activation, alternate metabolic routes, increased metabolic stability). Compounds with over-predicted activity may be subject to metabolic inactivation that can't be accounted for in the QSAR models. Further, we have demonstrated that $L_4$ has poor solubility that may account for its over-predicted activity.[233] Additionally, as can be seen from Figure 5.3, compound $L_4$ has extreme values of molecular weight and lipophilicity which may explain the inability of the generated QSAR models predict its activity. The trifluoromethyl groups on compounds $L_{49}$ and $L_6$, both with under-predicted activity, block metabolism at this site and also increase lipophilicity of these compounds. This leads to enhanced metabolic stability and facilitated passive diffusion across the lipophilic mycobacterial cell wall. These factors may have resulted in an improved MIC for these compounds which the QSAR model was not able to predict. Compounds $L_{64}$ and $L_{79}$ (CoMFA and CoMSIA outliers) both contain a metabolically labile carbamic ester functionality, cleavage of which could result in an active metabolite. This process may account for the under-predicted activity of these two compounds. Compound $L_{84}$ is unique in that it had a high residual when activity predictions were performed using the CoMSIA model (model 7), but residuals that did not result in outlier removal from any CoMFA model. As can be seen from Figure 5.6, for the most part the CoMFA and CoMSIA activity predictions were reasonably comparable; compounds $L_{84}$ and $L_{53}$ were the notable exceptions. The reason for the poor activity prediction of this compound by the CoMSIA model is not readily apparent.

### 5.3.4 Region Focusing

One method of 3D-QSAR optimization is known as region focusing.[254] It involves giving additional weight to the lattice points in a given CoMFA region to increase the contribution of those points in a further analysis. Region focusing is used to suppress PLS contributions from minor descriptors. The result is a new model with increased $q^2$ (cross-validated $r^2$), tighter grid spacing, and greater stability at a higher number of components. In this study, discriminant power was used to weight the lattice points by their contribution to the original model's components (see experimental methods). Figure 5.7 shows the CoMFA fields for one of the more active nitrofuran compounds before and after region focusing. As can be seen from the data for Model 23 in Table 5.4, the application of region focusing to Model 19 resulted in a significant improvement to the internal validity of the model, with small to negligible effect to the non-validated $r^2$ and test set activity predictions. Relative steric and electrostatic contributions were

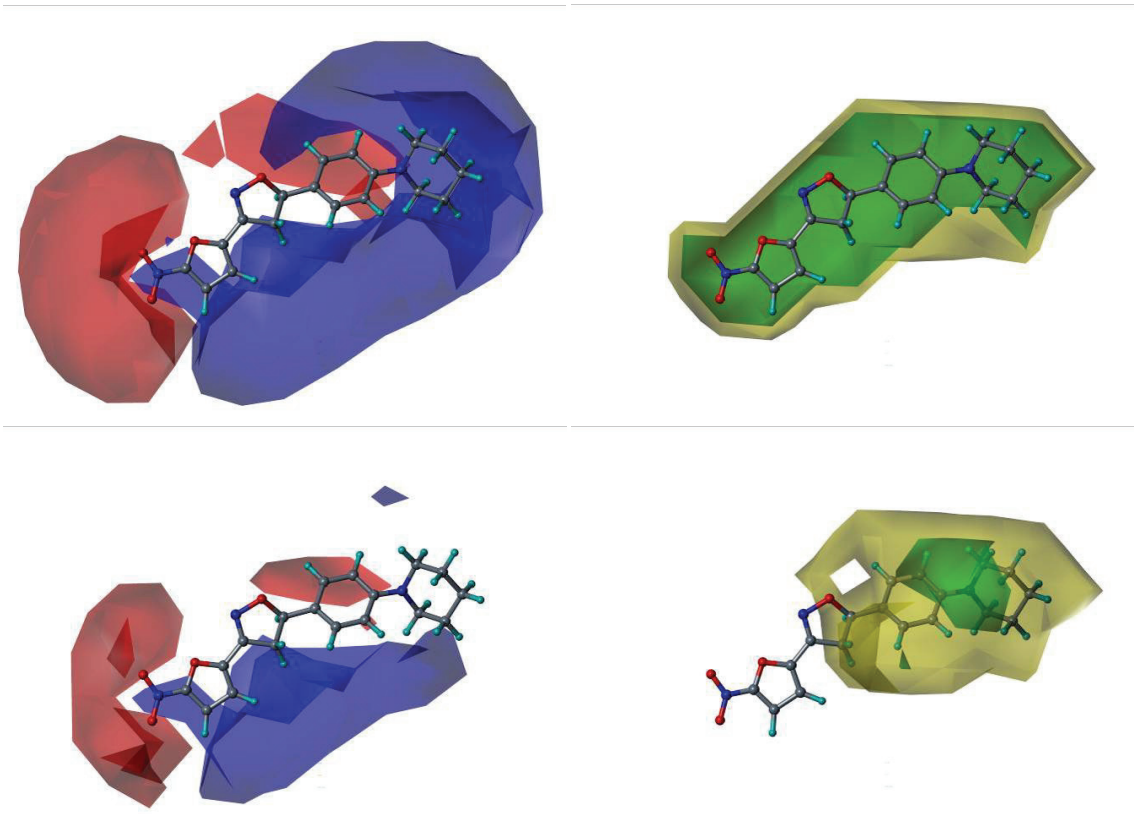


**Figure 5.7. QSAR Region Focusing**

The CoMFA field calculations are shown for $L_7$ before (upper) and after (lower) region focusing. Electrostatic fields (Left): Blue fields indicate electropositive groups favored, red fields indicate electronegative groups favored. Steric fields (Right): Green fields indicate steric bulk favored, yellow fields indicate steric bulk disfavored.

calculated from regression coefficients of the PLS models generated. Steric contributions played a larger role than electrostatic in the final model (model 23). The steric and electrostatic field contributions to the final model were 74% and 26%, respectively. Model 23 was selected as the best performing model in this 3D-QSAR study and will be used to predict the activity and guide future synthetic efforts on next generation nitrofuranyl compounds. Figure 5.8 graphically represents the biological activity predictions of Model 23. Figure 5.9 shows the CoMFA steric and electrostatic contour fields for the final model with the active compound, $L_{37}$, overlaid. Figure 5.10 displays the CoMSIA fields for our best performing CoMSIA model (model 22). The CoMFA fields indicate that the steric effects are mostly limited to the side chain, with clear areas seen where bulk is favored and disfavored. The CoMFA electrostatic fields show regions where positive and negative charges are favored on both the nitrofuran scaffold as well as the side chain. The blue field near the nitro group seems to indicate that compounds with less negative charge near the one of the nitro oxygens are favored; this is most likely due to the contribution of the aryl sulfone and aryl sulfoxide substitutions at this position in our training set. There is also a clear preference for a positively charged group at the terminal end of the side chain, which appears to correspond to basic amine groups at this position in several of the more active compounds in the training set. The CoMSIA fields (Figure 5.10) show steric regions and electrostatic fields that correlate well with what is seen in the CoMFA fields. Additional fields for hydrophobicity and H-bond donors and acceptors are shown; this information will be used for optimization of further generations of nitrofuran compounds.

### 5.3.5 Progressive Scrambling and Dependent Variable Scrambling

Cross-validation values must be interpreted with caution when building 3D-QSAR models with large training sets. This is because redundancy in the data sets can confuse the Leave-One-Out and Leave-Group-Out validation techniques.[252] The Progressive Scrambling method was developed to overcome this problem.[251-253] This method checks the sensitivity of the PLS model developed to small changes in the dependent variable. The values of $Q^2$, cSDEP, and $dq^2/dr^2_{yy}$' are returned and can aid in interpreting the predictivity of the model without the potentially confusing redundancy. The $Q^2$ statistic returned is an estimate of the predictivity of the model after removing the effects of redundancy. It is calculated by fitting the correlation of scrambled to unscrambled data ($r^2_{yy}$') to the cross-validated correlation coefficient ($q^2$) (calculated after each scrambling performed) using a 3rd order polynomial equation. The cSDEP statistic is an estimated crossvalidated standard error at a specific critical point (0.85 default used in this study) for $r^2_{yy}$', and is calculated from a 3rd order polynomial equation which fits the scrambled results. The slope of $q^2$ with respect to $r^2_{yy}$' is reported as $dq^2/dr^2_{yy}$', and is considered the critical statistic. It indicates to what extent the model changes with small changes to the dependent variable. In a stable model, $dq^2/dr^2_{yy}$' should not exceed 1.2 (ideally 1). This method was employed against the final model to verify the number of components used to build the model and to check the cross-validation against the possibility of such a redundancy in our training set. Table 5.5 lists the results of the

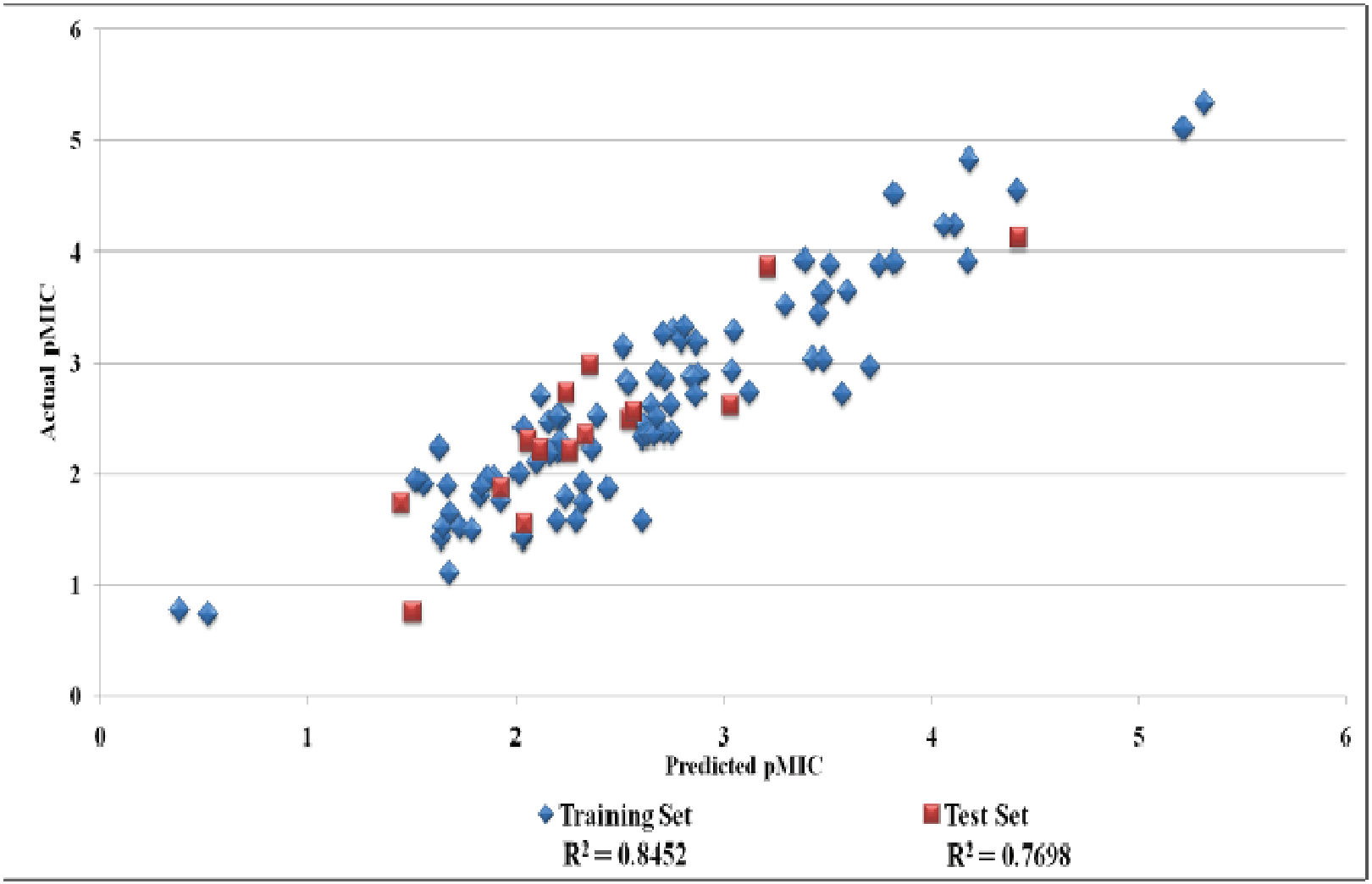



**Figure 5.8. Model 23 Results: Actual versus Predicted Activity**

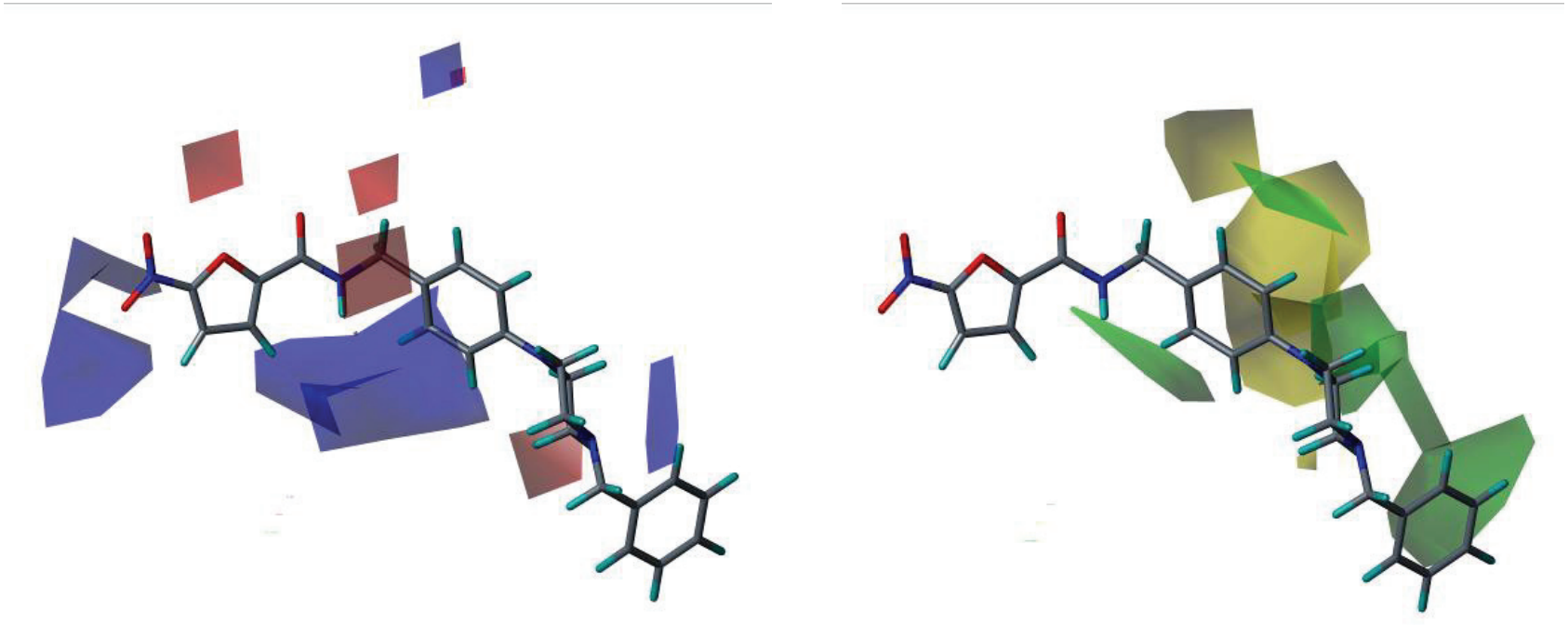


**Figure 5.9. CoMFA Field Contour Maps for Model 23 with Active Compound, L37**

Electrostatic fields (Left): Blue fields indicate electropositive groups favored, red fields indicate electronegative groups favored. Steric fields (Right): Green fields indicate steric bulk favored, yellow fields indicate steric bulk disfavored.

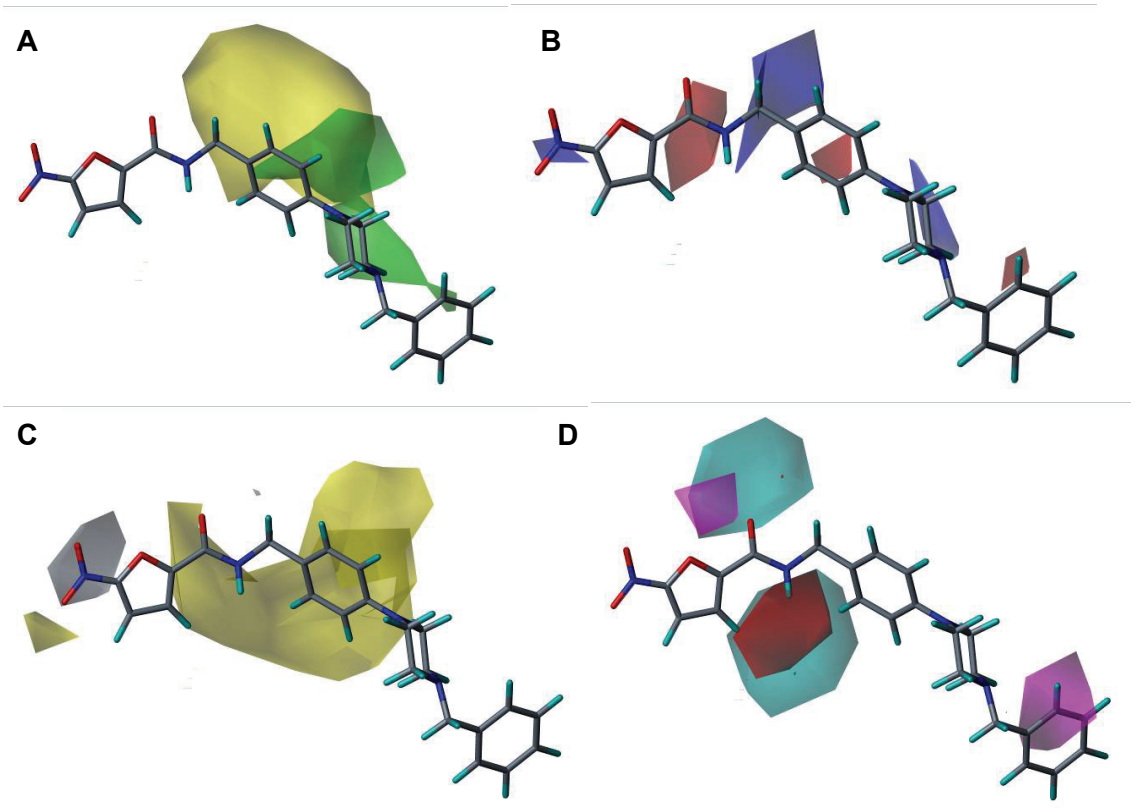



**Figure 5.10. CoMSIA Field Contour Maps for Model 22 with Active Compound, L37**

A. Steric Fields, Green indicates steric bulk favored, Yellow indicates bulk disfavored. B. Electrostatic fields, blue indicates positive charge favored, red indicates disfavored. C. Hydrophobic fields, Yellow indicates favored, gray indicates disfavored. D. H-bond donor and acceptor fields, Cyan indicates donor favored, Magenta indicates acceptor favored, and red indicates disfavored.[a]

a. H-bond donor disfavored fields were negligible at default energy values used for field generation and are not shown here.

**Table 5.5. Progressive Scrambling Results, Model 23**

| Components | $Q^2$ | cSDEP | $dq^{2'}/dr^2yy'$ |
|---|---|---|---|
| **2** | 0.337 | 0.776 | 0.13 |
| **3** | 0.387 | 0.750 | 0.52 |
| **4** | 0.430 | 0.726 | 0.78 |
| **5** | 0.432 | 0.728 | 1.15 |
| **6** | 0.381 | 0.763 | 1.47 |
| **7** | 0.424 | 0.741 | 1.48 |
| **8** | 0.393 | 0.766 | 1.55 |

progressive scrambling of Model 23. For a valid model, as additional components are added, values of $Q^2$ should be increasing while cSDEP is decreasing, the slope should fall near unity. While the value of the $Q^2$ statistic may seem low in comparison to the cross-validated $r^2$ ($q^2$) value, it must be noted that the introduced noise from scrambling renders this statistic very conservative. $Q^2$ values above 0.35 are reported to indicate that the original, unperturbed model is robust.[251] For Model 23, based on the progressive scrambling results, 5 components was the optimum number for use.

Another validation method that was employed in this study was Dependent Variable Scrambling (Y-scrambling). This method involves scrambling the dependent data in the training set and then building a PLS model using this scrambled data. The method is used to verify that the correlation in the original, unscrambled model is accurate and not a chance correlation. Ideally, the cross-validated $r^2$ ($q^2$) values returned from the scrambled PLS will be very low, even negatively correlated. Table 5.6 shows the results of the Y-scrambling test run against model 19. This model was chosen because model 23 was built by region focusing model 19, which was been built using unscrambled data. Therefore, Y-scrambling results against model 23 would not have been easily interpreted.

## 5.4 Summary

Using a series of nitrofuranyl compounds with known anti-tuberculosis activity, a predictive 3D-QSAR model has been developed. The effects of compound ionization, multiple alignments, and the incorporation of global molecular descriptors for lipophilicity, polar surface area, and steric bulk were investigated for their ability to improve QSAR model predictivity. Our expectation was that the addition of a lipophilicity descriptor (cLogP or LogD) and steric bulk descriptor could improve the model's predictivity by accounting for the cell entry contribution to the MIC of a given compound. We also theorized that polar surface area and ionization could model the effects of solubility.

**Table 5.6. Dependent Variable Scrambling Results, Model 19**

| Components | LOO $q^2$ | SEP |
|---|---|---|
| **1** | -0.260 | 1.210 |
| **2** | -0.546 | 1.349 |
| **3** | -0.498 | 1.335 |
| **4** | -0.833 | 1.486 |
| **5** | -0.863 | 1.507 |
| **6** | -0.827 | 1.501 |
| **7** | -0.765 | 1.485 |
| **8** | -0.791 | 1.505 |

Interestingly, the addition of molecular descriptors for lipophilicity, polar surface area, and steric bulk did little to improve the predictive ability of the model. While in most cases, the addition of the global molecular descriptors didn't weaken the models significantly, they did little to benefit them either. This may be due to the fact that most of the compounds in the training set had suitable physicochemical properties (cLogP 1-5) to penetrate the TB cell wall. As can be seen from Figure 5.3, although there is a clear trend of increasing activity with increased molecular weight, there is little correlation with cLogP in the range that our active compounds fall into. This is reflected in the QSAR models built in this study.

We noted above that the CoMFA steric field contribution of the final model (74%) greatly outweighed the electrostatic field contribution. As can be seen from the CoMFA fields shown in Figure 5.9 as well as the CoMSIA fields shown in Figure 5.10, the steric effects were isolated to the side chain while electrostatic effects were contributed from both the side chain and the nitrofuran scaffold. We believe this can be explained by the two processes discussed above, activation of the compounds by a nitro reducing enzyme (electrostatic effects, low steric contribution) and binding of the compound to its ultimate biological target (electrostatic and steric contribution). The CoMFA and CoMSIA fields clearly indicate regions of interest (both to avoid and to target) that will be used when performing CoMFA and/or CoMSIA guided activity predictions of nitrofurans for proposed synthesis and testing.

Another interesting result that we note is the improved performance of the QSAR models both in terms of internal validity and external (test set) predictivity when using alignment 2 versus alignment 1. In alignment 2, the side chains of the tertiary amide nitrofuran compounds adopted a conformation that was significantly different when compared to the unhindered nitrofurans and fell into a region not occupied by the unhindered compounds (see Figure 5.4, A). It is possible that this is reflecting the dual processes of compound activation and binding to the ultimate biomolecular target. While

it may seem from initial inspection of the CoMFA and CoMSIA fields in Figures 5.9 and 5.10 that these tertiary amide compounds contributed little to the final model, we point out that the test set included two such compounds whose activity was predicted with a fair degree of accuracy (within .5 pMIC units).

Further experiments are ongoing to investigate if our best performing models can be expanded to examine the nitroimidazole class of anti-tuberculosis agents. Preliminary evidence indicates that CoMFA model 23, discussed here, is suitable to predict MIC activity of these compounds as demonstrated by the reasonably accurate predictions of MIC's for PA824 (predicted 1.2 μg/mL, actual 0.5 μg/mL) and OPC67638 (predicted 0.0075 μg/mL, actual 0.006 μg/mL). This suggests that steric and electronic requirements for entry and nitro activation are shared by the nitrofuran and nitroimidazole anti-tuberculosis agents and are major contributors to this QSAR model.

The final model was optimized by outlier removal and region focusing and validated by a variety of methods; including cross-validation, progressive scrambling, and test set predictions. The model developed has high internal validity (cross-validated $r^2$ ($q^2$) above 0.5) and high predictive ability (test set $r^2$ above 0.7). It is being used to predict the anti-tuberculosis activity of proposed new compounds and to prioritize their synthesis by activity ranking. We believe this is an new important tool for the development of next generation nitrofuranyl and related nitroaromatic anti-tuberculosis agents.[233]

# CHAPTER 6. DISCUSSION AND CONCLUSIONS

## 6.1 General Dissertation Overview

This dissertation has presented my work on two research projects: the DHPS project focused on the identification of novel compounds that bind to the pterin subsite of dihydropteroate synthase and thereby inhibit enzyme activity; and the nitrofuran project focused on the advancement of a series of compounds with whole-cell activity against *M. tuberculosis*, both in terms of inhibitory activity and physical properties. These two projects afforded me the opportunity to use and evaluate a variety of computational tools to accomplish these research goals.

The availability of DHPS crystal structures with a variety of substrate and product analogs bound in the active site enabled the use of several structure-based drug design techniques, and were presented in chapters 2, 3 and 4. In Chapter 2, I presented my studies on the structure and function of DHPS and the mechanism of acquired resistance to sulfonamide agents. Using a series of molecular dynamics simulations, I was able to model the positions of loops that were either missing or incorrect from our crystal structures and visualize the locations of several residues that play key roles in both the reaction and resistance. Additionally, key insights into the role of the pterin subsite residues in ligand binding and the implications of these binding determinants in the design of pterin site inhibitors was discussed. An active site model was developed for use in subsequent high-throughput docking studies.

Chapter 3 presented the results of an extensive validation study of docking programs and scoring functions for use in high-throughput docking against the pterin binding site of DHPS. A variety of docking programs and scoring functions were thoroughly evaluated using several validation techniques including pose selection and scoring of a bound ligand, enrichment studies using receiver-operating characteristic curves, and a new metric designed specifically for this study, the SSLR statistic, which rewards both early enrichment and correct rank ordering by activity of known active compounds. In addition to selecting the best performing docking/scoring combination for use against DHPS, I was able to make general observations on the utility of the different validation metrics for use in validating against a single target.

Chapter 4 concluded my discussion of the DHPS virtual screening project by presenting the results of several large-scale, high-throughput molecular docking studies against the DHPS pterin site using fragment-based drug design concepts. These studies were built upon the work presented in the previous two chapters. Two successive rounds of docking were performed against the target using different screening techniques and the docking programs validated in Chapter 3. The first round of virtual screening used pharmacophore pre-filters, which enabled the screening of a very large number of compounds. Unfortunately, this led to “hit” compounds with

undesirable physico-chemical properties. The pharmacophore filtering step was removed for the second round of virtual screening, and resulted in far fewer compounds that we were able to screen, but with improved properties and novel scaffolds when compared to the first round. Ultimately, 15 fragment compounds were identified that had reasonable inhibitory activity against DHPS and they have been advanced to crystallography trials.

The second project discussed in this work was the nitrofuran project, which was presented in Chapter 5. This project used a series of compounds for which whole-cell activity against *M. tuberculosis* was known, but no structural target information was available. In this case, ligand-based techniques were employed, specifically 3D-QSAR studies, to generate a model which could be used to predict the activity of similar compounds that have not been synthesized or tested. The goal of these studies was to develop a model that can be used to advance the development of next generation nitrofuran compounds with improved physico-chemical properties and metabolic stability. The models developed in this study were generated using two new advanced techniques, Region Focusing and Progressive Scrambling, and were extensively validated. The combined use of these two methods allowed us to develop models with excellent predictivity.

## 6.2 Computational Medicinal Chemistry: A Diverse and Expanding Field

When employed by a skilled researcher with training and experience, the methods discussed below have the capability to identify active compounds which can be advanced to the clinic. This can be seen from the numerous examples of marketed drugs initially identified by these structure-based and ligand-based drug discovery methods (shown in Table 1.1). The key to success when applying these methods to drug discovery is the user. In addition to possessing expert skills in the use and function of the programs or algorithms which are being employed in the virtual screening study the Computational Medicinal Chemist should also be knowledgeable in four key complementary areas.

First, a working knowledge of organic chemistry is very important, even if the user is not performing any chemical reactions themselves. It is often necessary to filter out compounds from screening libraries that contain reactive or unstable functional groups. When making these filtering decisions, a background or working knowledge of organic chemistry is advantageous. Also, virtual libraries are often created using synthetic chemistry rules and building blocks available from commercial vendors. When designing virtual screening libraries, knowledge of synthetic organic chemistry is extremely helpful. Finally, the results of virtual screening studies are often used by synthetic chemists to either prioritize their synthesis projects or select compounds with high predicted activity for synthesis and testing. It is an advantage if the computational

chemist is able to provide suggestions to the organic chemists that are synthetically feasible, and a working knowledge of synthetic organic chemistry is helpful in facilitating discussion and interaction between the computational and synthetic organic chemist.

Second, knowledge of the basic principles of structural biology and the techniques used in the preparation of the atomic models that are used in structure-based drug design are absolutely essential. Knowing the limitations and uncertainties of an X-ray crystal or NMR structure being used for virtual screening is essential when preparing the structure for screening as well as when interpreting the results. An excellent paper published by Davis, et al. in 2003 highlighted important considerations the computational chemist must consider when using an X-ray crystal structure for a structure-based design project.[258] The biggest factor that must be understood is the uncertainty in the atomic positions of the structure being used and how to determine this uncertainty. Alternative side chain conformations and B factors can help the user identify areas of uncertainty in the atomic model. Molecular modeling environments are often inadequate for identifying these areas and expert interpretation of the structural files is necessary. Additionally, the laboratory methods and the experimental conditions used for determining these structures can have an effect on the structural model obtained, and this must be clearly understood by the computational chemist.

Third, an understanding of molecular biology techniques is also very important in terms of the experimental methods that are employed to measure the activity of any compounds being investigated. It has been said that high-throughput screening is only as good as the experimental assay being used in the screen. I would say that this is just as true when considering virtual screening because the “hit” compounds are typically intended to be tested for activity in an experimental assay. Knowledge of the assay conditions and limitations can be very important when identifying hit compounds and selecting compounds for testing because the physico-chemical properties of the compounds may be incompatible with the assay or assay conditions. In fact, *a priori* knowledge of the assay conditions and assay limitations can often influence the virtual screening parameters so as to select for compounds that are compatible with the assay. An understanding of other molecular biology techniques such as protein expression and isolation, gels, blotting, and arrays can also be useful to the computational medicinal chemist.

Finally, other areas that can be equally important include Anatomy and Physiology, Pharmacology, Pharmacokinetics and Pharmacodynamics, and Microbiology. Clearly, when attempting to discover compounds with biological activity for use in treating human disease, working knowledge of human pathophysiology and the molecular basis of drug action are very important. Traditionally, the pharmacokinetics and pharmacodynamics of lead compounds were optimized after identifying compounds with potent activity against the molecular target. However, due to the large number of clinical trial drug failures observed recently due to PK/PD issues and the exponential increase in the cost of bringing a drug to them market, these issues are

more frequently being addressed at earlier stages of drug discovery, even during hit and lead identification. There are now multiple algorithms and programs that are being used with increasing frequency to identify and eliminate compounds during the initial screening which are predicted to have unfavorable ADME or toxicological properties.[259] An understanding of these programs and properties is obviously very important.

## 6.3 Discussion of Methods

The studies discussed in this dissertation describe the use of a variety of molecular modeling and computer-aided drug design techniques, ranging from ligand-based methods of activity prediction to structure-based docking methods for hit compound identification. Every technique discussed in Chapter 1 has been used in these studies to some degree. These methods can be powerful tools to aid in the discovery, design and development of novel therapeutic agents, but it must be remembered that they are not without their limitations. Their full potential is only realized when their use is supported by other experimental methods, such as structural biology, molecular biology, microbiology, and organic synthesis. The studies presented here would not have been possible without contributions from researchers in each of these areas in multi-disciplinary, collaborative drug discovery projects. In Chapter 1, I introduced the virtual screening methods and tools used in these studies and discussed their theory and application. In this section, I present a critical assessment of the same by addressing their strengths and weaknesses in terms of their performance in these studies and discuss their future potential in more general terms.

### 6.3.1 Molecular Dynamics Simulations

When applied by a skilled user, molecular dynamics simulations can be employed to complement crystal studies by visualizing the motions of the atoms or residues of a biomolecular system. A crystal structure is a snapshot of a system, which may or may not represent the biomolecule in its native or active state. Simulations, on the other hand, can be used to visualize small scale movements such as loop movements, ligand binding, and possibly transition states. However, MD is not without limitations, and the studies discussed in Chapter 2 have demonstrated several that merit discussion. First and foremost, it must be remembered that MD simulations are just that: simulations. The methods used to obtain the energies and atomic positions are often approximations of approximations, and the results obtained from these studies must always be interpreted with this in mind and with a fair amount of caution.

MD simulations are time consuming and computationally expensive processes. In order to visualize loop movements (one of the goals of our simulations), an MD simulation must extend into the nanosecond range. Even when using time-saving shortcuts such as the SHAKE algorithm, energy cut-offs, PME electrostatic calculations, periodic boundary conditions, and implicit solvation, a nanosecond simulation can take a

week or more to run across multiple processors. A more rigorous simulation using explicit solvation can take up to several weeks. Additionally, in order to validate the model and results, it is often necessary to run multiple simulations to demonstrate the reproducibility of the results obtained. A project such as this can easily extend into months of work, as was the case in these studies. Visualization of large scale protein movements, protein folding, and rare events would require simulations into the millisecond range, and this is not possible with the current technology and computing power available.

Perhaps the most important consideration when setting up MD simulations is the parameterization of the non-native residues (ligands, cofactors, etc) present in the model being investigated. Most of the commonly used MD packages, such as the AMBER suite used in our studies, have developed parameters for the most common residues seen in biomolecules (amino acids, carbohydrates, nucleic acids, and even some commonly seen cofactors), and these have been extensively tested and validated. However, parameters (energy force constants and reference values for bond lengths, angles, torsions, etc.), must be developed, tested, and validated for any ligand or co-factor for which these parameters are not available, as was the case in our studies. We utilized the program Antechamber and the General Amber Force Field (GAFF) to derive parameters for non-native residues. Although this process is quick and usually reliable, it does not remove the need for testing and validating the derived parameters. This was the case for the parameters for our pyrophosphate ligand, which required extensive modification from those suggested by Antechamber before they could be used in our simulations to obtain reliable results. The main point here is that with molecular simulations, in order to obtain reliable results, both effort and skill are required to prepare the system prior to running any simulations. The old computational adage applies: garbage in, garbage out.

Finally, it should be mentioned that although the simulation methods employed in these studies are useful for visualizing small-scale protein movements and positions of ligands and side chains during binding, the force field methods used to obtain the energies are unable to show reactions, catalysis, or any process involving the flow of electrons. New methods including polarizable force fields and hybrid quantum mechanics/molecular modeling (QM/MM) methods have been recently reported that hope to address some of these deficiencies, but the methods still require extensive development and validation. Notwithstanding the limitations mentioned above, MD is a powerful tool that can be used very effectively in projects such as the one reported in this work, if one has a clear understanding of the limitations of the methods and keeps these limitations in mind when interpreting their results.

### 6.3.2 Structure-Based Drug Design

In a virtual screening project it is a definite advantage to know the structure of the biomolecular target being investigated. An X-ray, NMR, or even a homology modeled

structure makes a variety of structure-based design tools available to the researcher, most importantly docking and scoring techniques. Although these methods have demonstrated utility in the identification of lead compounds (as discussed in Chapter 1), they are not without their limitations. An understanding of these limitations is important to the successful application of these methods in a drug discovery project.

An obvious advantage of using docking and scoring as a lead identification method is the larger number of compounds that can be screened when compared to traditional high-throughput screening. Additionally, there is an added advantage in terms of both cost and time savings. However, expert decision making is necessary for both the preparation of the receptor for docking and the screening library which will be docked. Decisions regarding protonation states of ligands and side chains, charge calculation methods, and initial conformations must be made by the user prior to beginning the screen. These decisions typically require significant expertise on the part of the user with the program or programs that are being used to perform the docking run. A very large number of docking programs are now available for use today and it is unreasonable to expect a computational medicinal chemist to have achieved a high degree of expertise with more than a small number. Unfortunately, as has been demonstrated by several studies published to date, the performance of the docking program being used is at least partly dependent on the level of familiarity that the user has with the program.[185] In many docking program validation studies, the developers of a docking program are able to achieve significantly greater enrichment rates over other users, even when using identical program versions, biological targets, and screening libraries.

As mentioned in Chapter 3, not all docking programs and scoring functions will perform equally well when used to screen against a given target, regardless of the level of skill the user has with the programs. This is because there is a dependence on the nature of the binding site on the performance of the docking programs and scoring functions.[196] This can be attributed primarily to the functional form and parameterization of the scoring functions, with some performing better against polar active sites and others performing well against lipophilic sites. A clear understanding of these limitations is essential when selecting the appropriate docking and scoring functions for use against the screening target. In the absence of a thorough validation study as was described in Chapter 3, the computational chemist should make every effort to select a docking/scoring function combination with a proven record against the class of receptor into which they are docking.

Another limitation with structure-based virtual screening is the scoring functions themselves. A docking score is essentially an approximation of the binding affinity of a ligand for the receptor and should theoretically scale well with the experimentally determined binding affinity ($K_d$ or $IC_{50}$). However, although scoring functions have a demonstrated ability to identify active compounds from screening sets, they are nearly universal in poorly predicting the absolute binding affinity of active compounds. This has

direct ramifications on the selection of compounds from a virtual screen for testing. Compounds are often selected in one of two ways, either by taking the top n% of scored compounds or by taking every compound scoring above a certain cut-off. Either of these methods is acceptable, however the value selected (top n% or score cut-off) should be knowledge-based, that is based upon the results of a carefully designed validation study of the docking/scoring combination that was used against the target. A recommendation would be to select the n% value based on enrichment factor studies such that 80 to 90% of the active compounds were selected at the given n% value. Correspondingly, the score cut-off can be selected based upon receiver-operating characteristic curves by selecting the score above which 80 to 90% of the active compounds were identified. Beyond this, however, a testing priority for compounds based upon their actual score would be irrelevant.

Because of these limitations, it is very important to perform extensive validation of the docking/scoring function to be used in a virtual screening project. Additionally, a Medicinal Chemist who is performing molecular docking and scoring as part of their research should familiarize themselves with the functional form and parameterization of the scoring functions that are available for their use, and they should make every effort to obtain expertise with at least one docking program from each class (incremental construction, Monte Carlo, Genetic Algorithm, Tabu Search, etc.).

A final issue that deserves mention is compound procurement. Although not a direct limitation of a virtual screening project, the procurement of the "hit" compounds for testing is definitely a factor that requires consideration, as this process can require considerable expenditure of time and resources. In Chapter 4, I discussed the acquisition failure rates from the two rounds of virtual screening. Noticeable improvements to the failure rate were obtained when we selected only databases from reliable vendors, as determined from our experiences in the first round of virtual screening. In the absence of favorable experience otherwise, I would recommend limiting database screening to U.S. suppliers due to the high cost of shipping from foreign countries, customs issues, and questions of compound purity. In fact, although not specifically addressed in the research presented here, it is highly recommended to perform quality control analyses on all compounds ordered from any vendor. In the absence of some independent rating system for chemical suppliers (which does not exist to my knowledge), following these recommendations may help to alleviate some of the frustrations that were experienced in our studies.

### 6.3.3 Ligand-Based Drug Design

A recent article by Johnson, et al. entitled "The trouble with QSAR (or how I learned to stop worrying and embrace fallacy)" proposed that QSAR has not met the expectations for predicting biological activity.[260] The authors suggested that chance correlation, incorrect functional forms, and model overtraining have contributed to this problem and attributed the *Cum Hoc, Ergo Propter Hoc* fallacy (with this, therefore

because of this) to the poor prediction seen with many QSAR models.[260] They conclude that the manner in which QSAR is applied is more responsible for its lack of success than any other cause. The authors state their case convincingly and their paper correctly identifies issues and limitations with QSAR that must be addressed before any developed QSAR model can be used to make reliable activity predictions. However, I do believe that it is possible to develop and utilize a predictive QSAR model in a ligand-based drug design study if one understands the limitations mentioned above, takes them into account when developing the QSAR model, and extensively validates the model.

A question that I feel should be answered here is: Is there an advantage of using a QSAR model in drug discovery over general structure-activity observations? In other words, is QSAR a lesson in the obvious? In the case where a QSAR model tells you what you already know, what it the utility? The answer to these questions is that QSAR models do often tell us what we already know, a charged group favored in this position or steric bulk disfavored in that position. However, the true utility of QSAR models is in virtual screening. Knowing that a charged group is favored in this position and steric bulk disfavored in that is not helpful when predicting activities of a large number of compounds as a researcher cannot possibly be expected to visually inspect every compound for favored or disfavored groups within a reasonable period of time. However, a validated QSAR model can be used to screen very large libraries (of compounds covered by the physico-chemical space of the QSAR model) quickly and efficiently, providing the researcher a much more manageable number of "hit" compounds that can be visually inspected and ordered or synthesized.

## 6.4 Overall Themes ("The Big Picture")

Throughout the research presented in this thesis, there have been several recurring "themes" which deserve special attention. In this section I present a brief discussion of what I believe to be the two most important overall themes of the research I have conducted and the implications in virtual screening generally and with respect to the research projects discussed in this thesis.

### 6.4.1 Method Validation

The studies presented in this thesis predominately dealt with the use of computers, programs, and algorithms in the discovery of compounds with activity against our targets. I have implied above that these programs are essentially only as good as the researcher who is using them and that, in addition to expertise with the individual programs, a competent researcher should possess skills or a knowledge base in several fields contributing to drug discovery. In the previous section I highlighted and briefly discussed the multiple weaknesses of the methods used in these studies. Considering the expertise required for using these programs and their inherent weaknesses, the

question may be asked: How can one rely on the results obtained from the application of these methods?

The answer is perhaps the most important overall theme of this research: Validation. Each of the methods used in these studies, molecular dynamics, docking and scoring, and QSAR must be extensively validated against the target or system of interest before any results from their application can be reliable interpreted. It is not enough that the programs have been validated by their developers in a general sense; their use by a specific user against a specific target must also be validated. Chapter 3 presented an extensive validation of the docking and scoring methods we used in the virtual screening work presented in Chapter 4. The results presented in Chapter 3 highlight a key point that must be stressed here. The results obtained by the computational methods employed here are dependent on three factors: the expertise and training of the user, the abilities and limitations of the programs themselves, and the targets or systems against which they are being employed. This is demonstrated in Chapter 3 by the poor performance of several docking programs such as FlexX and GOLD when compared to others, although each of these programs has performed exceptionally well in other published studies. The question may be asked, when validating using the methods employed in these studies, which of the factors mentioned above is being investigated, the human factor, the programming factor, or the target factor? In many ways, the answer is all three. This of course has potential implications if the person or persons performing the validation studies are not the same who will be using the validated programs in the performance of the virtual screening, for example. Fortunately this was not the case in these studies.

Each of the computational techniques and programs used in these studies was validated using a variety of methods specific to the program or technique being utilized. For molecular dynamics simulations, validation methods include testing parameters by attempting to reproduce known experimental data such as thermodynamic properties, binding or conformational energies. Additionally, it is often necessary to run multiple simulations, often from different starting conformations, to investigate whether final structures and positions obtained can be reproduced. The results of a molecular dynamics simulation are questionable if they are not reproducible. Chapter 3 discussed the different validation methods currently employed for docking and scoring functions, and we have even developed a new validation method which we hope will be well received by the modeling community. The QSAR models developed in Chapter 4 were extensively validated by a variety of internal and external methods during and after their development.

### 6.4.2 Filtering and Compound Selection

The second general theme of this research that deserves special discussion is the selection of compounds for screening and testing. This is actually two separate but somewhat related areas, whose application directly influences the results obtained from

a virtual screen. In virtual screening, filters are frequently applied either before or after docking and scoring to minimize the computational expense and time required or the number of compounds required for testing. In our studies, we have used pharmacophore filtering as a docking pre-filter and cluster analysis as a docking post-filter. Our results, as discussed in Chapter 4, highlight some important issues with the application of filtering and compound selection that will be addressed in this section.

There are a variety of filters that can be employed for pre-filtering prior to performing a docking study, ranging from simple 2D filters for molecular weight or other physico-chemical properties, to advanced pharmacophore or SAR filters, such as those employed in our studies. In the case of pharmacophore filtering, the filter can be created from the structures of known active compounds or from key binding features that are known in the active site. These are known as ligand-based or receptor-based queries. In our case, we built a ligand-based query using the structural features of several compounds with DHPS inhibitory activity known to bind to the pterin site (crystal structures available). The results from the docking performed using this pharmacophore filter highlighted an important issue. Namely, the nature of the hit compounds from a virtual screen using this approach are dependent on how rigid the pharmacophore was that was employed as a pre-filter. A receptor-based pharmacophore filter would theoretically not be as rigid a pharmacophore, and could potentially enable more diversity in the hit compounds.

An alternative method would be to employ post-docking filters to minimize the number of compounds being sent for experimental assay studies. The number of compounds being docked and scored as well as the number of compounds being assayed are dependent on the resources available to a given research group. In our case, we decreased the number of compounds we docked by an order of magnitude and then applied cluster analysis to the results to select a manageable number of compounds for testing. Cluster analysis is one method of post-docking filtering that ideally selects a small number of compounds with maximum diversity, with the hope that the chemical space of the high scoring compounds is being adequately covered. The caveat with this method is that there is a greater chance of missing an active compound within the same bins. Although the testing from our second round of virtual screening is ongoing, it is my recommendation that a final step be added once more activity data on the compounds sent for testing becomes available. I suggest that, once a compound has been shown to have activity above our cut-off, then more compounds from the corresponding bin that compound was selected from should be procured and tested.

Very often in virtual screening projects, simple 2D filters are applied such as the Rule of 5 and Rule of 3 filters discussed in Chapter 1 and reactive or cytotoxic functional group filters such as those employed by the ZINC database curators (listed in Appendix D.1). These filters can be very useful in removing compounds that are not considered “drug-like” or “fragment-like”, have undesirable lipophilicity or electrostatic properties, or may react and interfere with experimental assays. However, it must be considered that

applying these filters may remove compounds with good activity from consideration. I will highlight this observation with two examples. First, when screening compounds for use against bacterial, fungal, or viral targets the use of "drug-like" filters is probably not appropriate (as mentioned in Section 1.5.3), because many marketed agents in these classes far exceed the cut-off values. Second, hits from a screen of fragment compounds are likely to be moved forward to an organic synthesis optimization project, and it may not be desirable to remove all compounds with reactive functional groups, as those groups may be advantageous for future synthesis and lead optimization. Obviously, the decision of what filters to employ and at what point to employ them will have to be made based upon a given projects goals and target.

Unless one has unlimited computational and experimental resources, which is rarely the case, filtering and compounds selection methods will almost always have to be employed in a high-throughput virtual screening project. Our studies have shown us that the filtering and selection rules employed in a virtual screening study are nearly as important as the docking and scoring functions in terms of the quality of the lead compounds obtained.

## 6.5 Future Directions

So, where do we go from here? Although my contributions to both the DHPS project and the Nitrofuran project are nearly completed, there are several avenues that remain to be explored, some revealed by the work presented here. Because both projects are still active, I present here some possible future directions that may be explored by DHPS and Nitrofuran researchers.

### 6.5.1 DHPS Project

The next logical step of this structure-based drug design project is the generation of co-crystal structures with the hit compounds from our VS studies that showed activity in our enzyme assay. In addition to validating our VS methods, a co-crystal structure with one or more of the hit fragments bound into the pterin site could be used by our synthetic chemistry group to generate more potent binding agents based upon the observed interactions in the structure and would also feed back into the modeling project to be used in refining the VS for future rounds.

Of course, virtual screening against the DHPS enzyme is far from complete. Although I have completed two extensive rounds of screening against *B. anthracis*, the DHPS project is also funded to investigate the enzyme from *F. tularensis* and *Y. pestis*. Crystal structures from these bacterial species were not available for my studies and I focused solely on *B. anthracis*. However, they should be solved very soon and investigation of these two enzymes in a manner similar to my investigations of *B. anthracis* DHPS is a logical next step. A very important prerequisite to high-throughput

docking against these enzymes is the development of a suitable assay for testing hit compounds. Extensive validation of the enzyme assay we are currently using, not only for *F. tularensis* and *Y. pestis*, but also for *B. anthracis* is necessary.

When we were validating docking programs and scoring functions for use in screening against DHPS, we investigated 5 commonly used docking programs and 9 scoring functions, all available to our group through University of Tennessee licenses. There were several notable exceptions missing from the docking programs that were investigated, and a closer examination of these programs for use in docking against DHPS would be useful. Two specific examples are the programs AutoDock and FRED, which are two of the most frequently used docking programs today for virtual screening studies (Figure 1.4). This is likely due to their free access to academic researchers, although both programs have also performed very well in validation studies. I was not able to investigate these programs due to "red tape" issues, but I feel that it is important to take a look at both AutoDock and FRED's performance against DHPS in the future. Additionally, two general classes of docking programs were noticeably missing from our validation studies, Monte Carlo based programs and Tabu Search based programs. The These were not available to our group at the time the studies were performed, but it would be very interesting to see how they compare with the programs that we did validate.

The "high-throughput" docking studies presented here used a Linux workstation with the docking jobs run in parallel across 4 processors (the most our University of Tennessee licenses permit). This enabled us to dock approximately 25,000 compounds per day. Our collaborators at St. Jude Children's Research Hospital have a 280-node Linux cluster and unlimited processor licenses available for their use. Using these resources, we could theoretically dock the entire ZINC database collection, without any pre-filtering, in a matter of days or weeks. Reports of this type of "ultra high-throughput docking" are scarce and this is definitely something that merits investigation by the group. Of course, the important issues of validation and compound selection for testing would have to be addressed. The top 1% of 5 million compounds is 50,000 compounds!

In our docking validation study, we investigated nine different scoring functions for use in screening against the DHPS pterin site. At least one scoring function was present from each major class of scoring functions: force-field based, knowledge-based, and empirical functions. Of interest to me personally, and perhaps of utility to future virtual screening studies, is the use of solvation-based scoring functions. At the most, each of the energy functions investigated in these studies very generally approximates the effects of solvation/desolvation on ligand binding, usually through the incorporation of a protein-ligand desolvation energetic penalty in the scoring function. There are continuum based approaches that represent an intermediate approach to the incorporation of solvation effects in scoring, such as GB/SA scoring functions (not investigated in these studies). The reason that a full solvation-based scoring approach, such as PB/SA scoring, was avoided here was due to the dramatic increase in time and

computing power that would have been required. However, new techniques and increased ability to parallelize such docking and scoring jobs are now available. It would be of interest to this study to use the Linux cluster at SJCRH with the latest version of the program DOCK, which is now highly parallelizable using MPI software freely available, and the PB/SA or AMBER GB/SA scoring function (or both) scoring function that comes with the DOCK package.

With respect to the Molecular Dynamics simulations performed in this study, I do not feel that we learned as much as I had hoped. After analyzing the results of multiple simulations, it seemed that there were more questions than answers. Although, a few interesting observations were noted and theories proposed as to the nature of certain key interactions and resistance site mutations, ultimately our simulations failed to answer any of our research questions clearly. Should any future researcher in this area decide to take another look at DHPS using Molecular Simulations I would recommend the following: 1. QM calculations to be used for the development of non-standard residue parameters. This will avoid the time-consuming trial and error that was experienced in these studies. 2. Extended duration simulations using product, substrate, and transition-state analogs using advanced MD techniques such as umbrella sampling[261], replica exchange[262], or multicanonical ensembles[263], may enable us to visualize loop movements that were not seen with the 4 nanosecond timescale used in these studies. 3. The use of new Quantum Dynamics or QM/MM simulations to attempt to more accurately study the transition state and perhaps definitively answer the question of whether the reaction proceeds via an Sn1 or Sn2 mechanism.

Lastly, there are two important areas that I think should be investigated closely in future studies. The first is the conserved water binding site which falls deep within the DHPS active site, near the pterin binding site. Although we avoided targeting this site in the docking studies presented here due to its conserved nature, this water site has potential to be displaced by a small molecule moiety to yield slow-, tight-binding inhibitors of DHPS which could lead to broad-spectrum antibiotics. The displacement of a structural water has been reported to be one mechanism of slow-, tight-binding and could theoretically lead to inhibitors that are functionally equivalent to covalent, irreversible inhibitors with an extended pharmacokinetic half-life of days.[264] The implications in antimicrobial drug design are obvious. The second area is the phosphate binding site, which was also avoided in our docking studies. There are two reasons for my belief that this is an area that should also be investigated in future virtual screens or synthetic efforts. First, in our molecular dynamics simulations we noted a dependence of an anionic group in this site for stabilization of the pterin product and substrate ligands in the active site and a rapid ejection of products and substrates when the negatively charged group was absent. Second, several of our most active hit compounds from the virtual screening studies and other preliminary studies contain a negatively charged group which could theoretically be falling near this anionic site. It is very likely that the anion stabilizes a key arginine in an extended position that is necessary for binding of

pterin substrates. Theoretically, inhibitory compounds containing a similar anionic group would have an increased binding affinity due to their ability to make similar contacts.

### 6.5.2 Nitrofuran Project

With the development and validation of a predictive QSAR model, the next obvious step is to utilize the model to predict activities of unknown compounds and to use this information to prioritize their synthesis and testing. I suggest that the next researcher who is assigned to assist with modeling on this project develop a series of virtual nitrofuran (and related compound) libraries based loosely upon the scaffold of the compounds in our collection with potent activity. I recommend that the modeler work closely with the synthetic group when preparing these libraries to facilitate the development of a library containing synthetically feasible compounds. As new compounds are tested and activities measured, the QSAR model can be expanded to cover the additional chemical space. Of course any new models developed will have to be extensively validated.

Finally, the Nitrofuran project would benefit from the application of one or more lead-hopping techniques to identify novel scaffolds. There are two methods that I would suggest be applied in this case: similarity searching and topomer searching, both of which are available through our Sybyl Molecular Modeling suite. The first program I recommend be employed is Surflex-Sim, which bases similarity on the training set molecules' shape, H-bonding, and electrostatic properties using molecular surfaces. The second is a new technique called topomer searching, which is an extremely fast tool for ligand-based VS and lead-hopping. It uses topomer fields and pharmacophore properties of the training compounds to screen for whole molecules, groups, or scaffolds. The application of either of these methods has the potential to take the nitrofuran project in exciting new directions.

## 6.6 Conclusions

I conclude this work by giving my best answer to the following question: What place do I believe virtual screening has in the future of drug design?

There are still several inadequacies with virtual screening methods that will need to be overcome before virtual screening can realize its full potential. A key weakness is the inability of any computational method to accurately predict absolute binding affinity. Although VS has advanced to the point where it can be used to reliable identify active compounds from screening libraries, the functions themselves have a long way to go before they can be used to predict the absolute binding affinity of the compounds with a reliable degree of certainty. This was demonstrated by our use of the new SSLR statistic in Chapter 4. If the scoring functions were getting the ordering even close to correct, the SSLR metric should theoretically have statistically outperformed the AU-

ROC method. The fact that the performance of these two methods was mostly indistinguishable reflects on the performance of the scoring functions rather than the utility of the SSLR method.

However, I still firmly believe that virtual screening as a lead identification method has strong future in drug discovery. In fact, as the techniques continue to be refined and improved and as computing power continues to exponentially increase, I believe that the deficiencies highlighted in this chapter will be resolved and that the role of virtual screening in drug discovery will continue to expand. Eventually, it may even replace traditional high-throughput screening as the gold standard of lead identification. Computational medicinal chemistry, while still an emerging field, will play an increasing role in the discovery of future clinical drug candidates, not just as a lead identification tool but also as a lead optimization tool. I foresee an increased demand for researchers with skill and training in these techniques.

## LIST OF REFERENCES


1. Hassall, C. H. Computer graphics as an aid to drug design. *Chem Brit* **1985,** 21, 39-46.

2. Quiocho, F. A.; Lipscomb, W. N. Carboxypeptidase A: a protein and an enzyme. *Adv Protein Chem* **1971,** 25, 1-78.

3. Cushman, D. W.; Cheung, H. S.; Sabo, E. F.; Ondetti, M. A. Design of potent competitive inhibitors of angiotensin-converting enzyme. Carboxyalkanoyl and mercaptoalkanoyl amino acids. *Biochemistry* **1977,** 16, 5484-91.

4. Patick, A. K.; Potts, K. E. Protease inhibitors as antiviral agents. *Clin Microbiol Rev* **1998,** 11, 614-27.

5. Abdel-Magid, A. F.; Maryanoff, C. A.; Mehrman, S. J. Synthesis of influenza neuraminidase inhibitors. *Curr Opin Drug Discov Devel* **2001,** 4, 776-91.

6. Summa, V.; Pace, P.; Petrocchi, A.; Laufer, R.; Cortese, R.; Hazuda, D. J.; Miller, M. D.; Schleif, W. A.; Vacca, J. P.; Young, S. D.; Rowley, M. Discovery of MK-0518 a novel, potent and selective HIV integrase inhibitor in phase III clinical trials. In *16th International AIDS Conference*, Toronto, Canada, **2006**.

7. Kawakami, Y.; Inoue, A.; Kawai, T.; Wakita, M.; Sugimoto, H.; Hopfinger, A. J. The rationale for E2020 as a potent acetylcholinesterase inhibitor. *Bioorg Med Chem* **1996,** 4, 1429-46.

8. Deininger, M. W.; Goldman, J. M.; Lydon, N.; Melo, J. V. The tyrosine kinase inhibitor CGP57148B selectively inhibits the growth of BCR-ABL-positive cells. *Blood* **1997,** 90, 3691-8.

9. Moyer, J. D.; Barbacci, E. G.; Iwata, K. K.; Arnold, L.; Boman, B.; Cunningham, A.; DiOrio, C.; Doty, J.; Morin, M. J.; Moyer, M. P.; Neveu, M.; Pollack, V. A.; Pustilnik, L. R.; Reynolds, M. M.; Sloan, D.; Theleman, A.; Miller, P. Induction of apoptosis and cell cycle arrest by CP-358,774, an inhibitor of epidermal growth factor receptor tyrosine kinase. *Cancer Res* **1997,** 57, 4838-48.

10. Hardy, L. W.; Malikayil, A. The impact of structure-guided drug design on clinical agents. *Current Drug Discovery* **2003**, 15-20.

11. Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J Med Chem* **2002,** 45, 2213-21.

12. Klebe, G. Virtual screening: Scope and limitations. In *Virtual Screening in Drug Discover*, Alvarez, J. C.; Shoichet, B. K., Eds. CRC Press, Taylor and Francis Group: Boca Raton, FL, **2005**; pp 1-24.

13. Rishton, G. M. Reactive compounds and in vitro false positives in HTS. *Drug Discov Today* **1997,** 2, 382-4.

14. McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J Med Chem* **2002,** 45, 1712-22.

15. Babaoglu, K.; Simeonov, A.; Irwin, J. J.; Nelson, M. E.; Feng, B.; Thomas, C. J.; Cancian, L.; Costi, M. P.; Maltby, D. A.; Jadhav, A.; Inglese, J.; Austin, C. P.; Shoichet, B. K. Comprehensive mechanistic analysis of hits from high-throughput and docking screens against beta-lactamase. *J Med Chem* **2008,** 51, 2502-11.

16. Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. "Lead hopping". Validation of topomer similarity as a superior predictor of similar biological activities. *J Med Chem* **2004,** 47, 6777-91.

17. Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening. *Angew Chem Int Ed Engl* **1999,** 38, 2894-6.

18. Jenkins, J. L.; Glick, M.; Davies, J. W. A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *J Med Chem* **2004,** 47, 6144-59.

19. Zhang, Q.; Muegge, I. Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. *J Med Chem* **2006,** 49, 1536-48.

20. Franke, L.; Schwarz, O.; Muller-Kuhrt, L.; Hoernig, C.; Fischer, L.; George, S.; Tanrikulu, Y.; Schneider, P.; Werz, O.; Steinhilber, D.; Schneider, G. Identification of natural-product-derived inhibitors of 5-lipoxygenase activity by ligand-based virtual screening. *J Med Chem* **2007,** 50, 2640-6.

21. Boehm, M.; Wu, T. Y.; Claussen, H.; Lemmen, C. Similarity searching and scaffold hopping in synthetically accessible combinatorial chemistry spaces. *J Med Chem* **2008,** 51, 2468-80.

22. Saeh, J. C.; Lyne, P. D.; Takasaki, B. K.; Cosgrove, D. A. Lead hopping using SVM and 3D pharmacophore fingerprints. *J Chem Inf Model* **2005,** 45, 1122-33.

23. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **1997,** 23, 3-25.

24. Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* **2002,** 45, 2615-23.

25. Leach, A. R. *Molecular Modelling: Principles and Applications*. 2nd ed.; Prentice Hall: London, England, **2001**; p 743.

26. Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared near neighbours. *IEEE Transactions in Computers* **1973,** C-22, 1025-34.

27. Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **1967,** 32, 241-54.

28. Downs, G. M.; Willett, P.; Fisanick, W. Similarity searching and clustering of chemical-structure databases using molecular property data. *J Chem Inf Comput Sci* **1994,** 34, 1094-1102.

29. Brown, R. D.; Martin, J. C. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J Chem Inf Comput Sci* **1996,** 36, 572-83.

30. Hansch, C. A quantitative approach to biochemical structure-acitivty relationships. *Accounts of Chemical Research* **1969,** 2, 232-9.

31. Smith, R. N.; Hansch, C.; Ames, M. M. Selection of a reference partitioning system for drug design work. *J Pharm Sci* **1975,** 64, 599-606.

32. Scherrer, R. A.; Howard, S. M. Use of distribution coefficients in quantitative structure-activity relationships. *J Med Chem* **1977,** 20, 53-8.

33. Hansch, C.; Leo, A.; Unger, S. H.; Kim, K. H.; Nikaitani, D.; Lien, E. J. "Aromatic" substituent constants for structure-activity correlations. *J Med Chem* **1973,** 16, 1207-16.

34. Stanton, D. T.; Dimitrov, S.; Grancharov, V.; Mekenyan, O. G. Charged partial surface area (CPSA) descriptors QSAR applications. *SAR QSAR Environ Res* **2002,** 13, 341-51.

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

36. Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J Med Chem* **1994,** 37, 4130-46.

37. Silverman, B. D.; Platt, D. E. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J Med Chem* **1996,** 39, 2129-40.

38. Ferguson, A. M.; Heritage, T.; Jonathon, P.; Pack, S. E.; Phillips, L.; Rogan, J.; Snaith, P. J. EVA: a new theoretically based molecular descriptor for use in QSAR/QSPR analysis. *J Comput Aided Mol Des* **1997,** 11, 143-52.

39. Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D QSAR study in a series of steroids. *J Comput Aided Mol Des* **1997,** 11, 79-92.

40. Doweyko, A. M. The hypothetical active site lattice. An approach to modelling active sites from data on inhibitor molecules. *J Med Chem* **1988,** 31, 1396-406.

41. Wold, S.; Albano, C.; Dunn III, W. J.; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjostrom, M. Multivariate data analysis in chemistry. In *CHEMOMETRICS: Mathematics and Statistics in Chemistry*, Kowalski, B., Ed. Reidel: Dordrecht, Netherlands, **1984**.

42. Andrea, T. A.; Kalayeh, H. Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. *J Med Chem* **1991,** 34, 2824-36.

43. King, R. D.; Muggleton, S.; Lewis, R. A.; Sternberg, M. J. Drug design by machine learning: the use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc Natl Acad Sci U S A* **1992,** 89, 11322-6.

44. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* **1977,** 112, 535-42.

45. Note: A SCOPUS query (excluding reviews) for 2007 citations using the original references for the top 20 docking programs shown in Figure 3 returns 847 journal articles.

46. Morris, G. M.; DGoodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **1998,** 19, 1639-62.

47. Wu, G.; Robertson, D. H.; Brooks, C. L., 3rd; Vieth, M. Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMm-based MD docking algorithm. *J Comput Chem* **2003,** 24, 1549-62.

48. Taylor, J. S.; Burnett, R. M. DARWIN: a program for docking flexible molecules. *Proteins* **2000,** 41, 173-91.

49. Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J Comput Chem* **1997,** 18, 1175-89.

50. Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y. P. Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J Med Chem* **2000,** 43, 401-8.

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

52. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **1996,** 261, 470-89.

53. Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *J Comput Aided Mol Des* **1994,** 8, 153-74.

54. McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003,** 68, 76-90.

55. Gabb, H. A.; Jackson, R. M.; Sternberg, M. J. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* **1997,** 272, 106-20.

56. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* **2004,** 47, 1739-49.

57. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **1997,** 267, 727-48.

58. Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem Biol* **1996,** 3, 449-62.

59. Abagyan, R.; Totrov, M.; Kuznetzov, D. ICM - a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J Comput Chem* **1994,** 15, 488-506.

60. Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J Mol Graph Model* **2003,** 21, 289-307.

61. Sobolev, V.; Wade, R. C.; Vriend, G.; Edelman, M. Molecular docking using surface complementarity. *Proteins* **1996,** 25, 120-9.

62. Bohm, H. J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J Comput Aided Mol Des* **1992,** 6, 61-78.

63. Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **1998,** 33, 367-82.

64. McMartin, C.; Bohacek, R. S. QXP: powerful, rapid computer algorithms for structure-based drug design. *J Comput Aided Mol Des* **1997,** 11, 333-44.

65. Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem* **2003,** 46, 499-511.

66. Burnham, J. F. Scopus database: a review. *Biomed Digit Libr* **2006,** 3, 1.

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

68. Mizutani, M. Y.; Tomioka, N.; Itai, A. Rational automatic search method for stable docking models of protein and ligand. *J Mol Biol* **1994,** 243, 310-26.

69. Trosset, J. Y.; Scheraga, H. A. Prodock: software package for protein modeling and docking. *J Comput Chem* **1999,** 20, 412-27.

70. Liu, M.; Wang, S. MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *J Comput Aided Mol Des* **1999,** 13, 435-51.

71. Clark, K. P. Flexible ligand docking without parameter adjustment across four ligand-receptor complexes. *J Comput Chem* **1995,** 16, 1210-26.

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

73. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calulations. *J Comp Chem* **1983,** 4, 187-217.

74. Kale, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. NAMD2: Greater scalability for parallel molecular dynamics. *Journal of Computational Physics* **1999,** 151, 283-312.

75. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. GROMACS: fast, flexible, and free. *J Comput Chem* **2005,** 26, 1701-18.

76. Leach, A. R. Energy minimization and related methods for exploring the energy surface. In *Molecular Modelling: Principles and Applications*, 2nd ed.; Pearson Education Limited: Harlow, England, **2001**; pp 253-302.

77. Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins* **1999,** 37, 228-41.

78. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins* **2003,** 52, 609-23.

79. Bohm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des* **1994,** 8, 243-56.

80. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des* **1997,** 11, 425-45.

81. Rognan, D.; Lauemoller, S. L.; Holm, A.; Buus, S.; Tschinke, V. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem* **1999,** 42, 4650-8.

82. Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem* **1999,** 42, 791-804.

83. Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* **2000,** 295, 337-56.

84. DeWhitte, R. S.; Shakhnovich, E. I. SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J Am Chem Soc* **1996,** 118, 11733-44.

85. Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J Mol Graph Model* **2002,** 20, 281-95.

86. Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des* **2002,** 16, 11-26.

87. Rashin, A. A.; Namboodiri, K. A simple method for the calculation of hydration enthalpies of polar molecules with arbitrary shapes. *J Phys Chem* **1987**, 916003-12.

88. Zou, X.; Sun, Y.; Kuntz, I. D. Inclusion of solvation in ligand binding free energy calculations using the genralized-born model. *J Am Chem Soc* **1999**, 1218033-43.

89. Majeux, N.; Scarsi, M.; Caflisch, A. Efficient electrostatic solvation model for protein-fragment docking. *Proteins* **2001,** 42, 256-68.

90. ZAP, OpenEye Scientific Software, www.eyesopen.com. Accessed April 1, 2008.

91. Zhang, T.; Koshland, D. E., Jr. Computational method for relative binding energies of enzyme-substrate complexes. *Protein Sci* **1996,** 5, 348-56.

92. Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* **1999,** 42, 5100-9.

93. Rush, T. S.; Manas, E. S.; Tawa, G. J.; Alvarez, J. C. Solvation-based scoring for high throughput docking. In *Virtual Screening in Drug Discovery*, Alvarez, J. C.; Shoichet, B. K., Eds. Taylor & Francis: Boca Raton, FL, **2005**; pp 249-77.

94. Apostolakis, J.; Pluckthun, A.; Caflisch, A. Docking small ligands in flexible binding sites. *J Comput Chem* **1998,** 19, 21-37.

95. Schnecke, V.; Swanson, C. A.; Getzoff, E. D.; Tainer, J. A.; Kuhn, L. A. Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. *Proteins* **1998,** 33, 74-87.

96. Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J Mol Biol* **1997,** 266, 424-40.

97. Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins* **2002,** 46, 34-40.

98. Oprea, T.; Bologa, C.; Olah, M. Compound selection for virtual screening. In *Virtual Screening in Drug Discovery*, Alvarez, J. C.; Shoichet, B. K., Eds. Taylor & Francis: Boca Raton, FL, **2005**; pp 89-106.

99. Bostrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J Mol Graph Model* **2003,** 21, 449-62.

100. Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of automatic three-dimensional model builders Using 639 X-ray Structures. *J Chem Inf Comput Sci* **1994,** 34, 1000-8.

101. Pearlman, R. S., "Concord," distributed by Tripos International, St. Louis, Missouri, 63144, USA.

102. Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comp Method.* **1990,** 3, 537-47.

103. Stewart, J. J. P. Optimization of parameters for semi-empirical methods. *J Comput Chem* **1989,** 10.

104. Del Re, G. A simple MO-LCAO method for the calculation of charge distributions in saturated organic molecules. *J Chem Soc* **1958**, 4031-40.

105. Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron* **1980,** 36, 3219-28.

106. Purcell, W. P.; Singer, J. A. A brief review and table of semiempirical parameters used in the Huckel Molecular Orbital method. *J Chem Eng Data* **1967,** 12, 235-46.

107. Berthod, H.; Pullman, A. Calculation of the structure of conjugated molecules. *J Chem Phys* **1965,** 62, 942-6.

108. Halgren, T. A. Maximally diagonal force constants in dependent angle-bending coordinates. 2. Implications for the desisng of empirical force fields. *J Am Chem Soc* **1990,** 112, 4710-23.

109. Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alogona, G.; Profeta, S.; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J Am Chem Soc* **1984,** 106, 765-84.

110. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* **1995,** 117, 5179-97.

111. MacKerell, A. D., Jr.; Brooks, B.; Brooks, C. L., 3rd; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. CHARMM: The energy function and its paramerterization with an overview of the program. In *The Encyclopedia of Computational Chemistry*, Schleyer, P. v. R., Ed. John Wiley & Sons: Chichester, **1998**; Vol. 1, pp 271-7.

112. Jorgensen, W. L.; Tirado-Rives, J. The OPLS force field for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc* **1988,** 110, 1657-66.

113. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* **1996,** 118, 11225-36.

114. Schuler, L. D.; Daura, X.; van Gunsteren, W. F. An improved GROMOS96 force field for aliphatic hydrocarbons in the conensed phase. *J Comput Chem* **2001,** 22, 1205-18.

115. Maple, J. R.; Hwang, M.-J.; Stockfisch, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T. Derivation of class II force fields. I. Methodology and quantum force field for the alkyl functional group and alkane molecules. *J Comput Chem* **1994,** 15, 161-82.

116. Allinger, N. L. Conformational analysis 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. *J Am Chem Soc* **1977,** 99, 8127-34.

117. Allinger, N. L.; Li, F.; Yan, L. Molecular mechanics. The MM3 force field for alkenes. *J Comput Chem* **1990,** 11, 848-67.

118. Allinger, N. L.; Li, F.; Yan, L. Molecular mechanics (MM3) calculations on conjugated hydrocarbons. *J Comput Chem* **1990,** 11, 868-95.

119. Allinger, N. L.; Chen, K.; Katzenelenbogen, J. A.; Wilson, S. R.; Anstead, G. M. Hyperconjugative effects on carbon-carbon bond lengths in molecular mechanics (MM4). *J Comput Chem* **1996,** 17, 747-55.

120. Allinger, N. L.; Chen, K.; Lii, J.-H. An improved force field (MM4) for saturated hydrocarbons. *J Comput Chem* **1996,** 17, 642-68.

121. Plimpton, S. J. Fast parallel algorithms for short-range molecular dynamics. *J Comp Phys* **1995,** 117, 1-19.

122. Verlet, L. Computer 'Experiments' on Classical Fluids. I. Thermodynamic properties of Lennard-Jones molecules. *Physical Review* **1967,** 159, 98-103.

123. Hockney, R. W. The potential calculation and some applications. *Method in Computational Physics* **1970,** 9, 136-211.

124. Swope, W. C.; Anderson, H. C.; Berens, P. H.; Wilson, K. R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: application to small water clusters. *Journal of Chemical Physics* **1982,** 76, 637-49.

125. Ryckaert, J. P.; Cicotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comp Phys* **1977,** 23, 327-41.

126. Ewald, P. Due Berechnung optischer udn elektrostatischer gitterpotentiale. *annalen der physik* **1921,** 64, 253-87.

127. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **1983,** 79, 926-35.

128. Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The design of leadlike combinatorial libraries. *Angew Chem Int Ed Engl* **1999,** 38, 3743-8.

129. Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. Fragment-based lead discovery. *Nat Rev Drug Discov* **2004,** 3, 660-72.

130. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'rule of three' for fragment-based lead discovery? *Drug Discov Today* **2003,** 8, 876-7.

131. Hann, M. M.; Leach, A. R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J Chem Inf Comput Sci* **2001,** 41, 856-64.

132. Leach, A. R.; Hann, M. M.; Burrows, J. N.; Griffen, E. J. Fragment screening: an introduction. In *Structure-Based Drug Discovery*, Hubbard, R. E., Ed. RSC: Cambridge, **2006**; pp 430-46.

133. Fattori, D. Molecular recognition: the fragment approach in lead generation. *Drug Discov Today* **2004,** 9, 229-38.

134. Nienaber, V. L.; Richardson, P. L.; Klighofer, V.; Bouska, J. J.; Giranda, V. L.; Greer, J. Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nat Biotechnol* **2000,** 18, 1105-8.

135. Carr, R.; Jhoti, H. Structure-based screening of low-affinity compounds. *Drug Discov Today* **2002,** 7, 522-7.

136. Moy, F. J.; Haraki, K.; Mobilio, D.; Walker, G.; Powers, R.; Tabei, K.; Tong, H.; Siegel, M. M. MS/NMR: a structure-based approach for discovering protein ligands and for drug design by coupling size exclusion chromatography, mass spectrometry, and nuclear magnetic resonance spectroscopy. *Anal Chem* **2001,** 73, 571-81.

137. Erlanson, D. A.; McDowell, R. S.; O'Brien, T. Fragment-based drug discovery. *J Med Chem* **2004,** 47, 3463-82.

138. Ciulli, A.; Williams, G.; Smith, A. G.; Blundell, T. L.; Abell, C. Probing hot spots at protein-ligand binding sites: a fragment-based approach using biophysical methods. *J Med Chem* **2006,** 49, 4992-5000.

139. Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discov Today* **2004,** 9, 430-1.

140. Davies, T. G.; van Montfort, R. L.; Williams, G.; Jhoti, H. Pyramid: an integrated platform for fragment-based drug discovery. In *Fragment-based Approaches in Drug Discovery*, Jahnke, W.; Erlanson, D. A., Eds. Wiley-VCH: Weinheim, **2006**; pp 193-214.

141. Card, G. L.; Blasdel, L.; England, B. P.; Zhang, C.; Suzuki, Y.; Gillette, S.; Fong, D.; Ibrahim, P. N.; Artis, D. R.; Bollag, G.; Milburn, M. V.; Kim, S. H.; Schlessinger, J.; Zhang, K. Y. A family of phosphodiesterase inhibitors discovered by cocrystallography and scaffold-based drug design. *Nat Biotechnol* **2005,** 23, 201-7.

142. Mayer, M.; Meyer, B. Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angew Chem Int Ed* **1999,** 38, 1784-88.

143. Dalvit, C.; Fogliatto, G.; Stewart, A.; Veronesi, M.; Stockman, B. WaterLOGSY as a method for primary NMR screening: practical aspects and range of applicability. *J Biomol NMR* **2001,** 21, 349-59.

144. Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **1996,** 274, 1531-4.

145. Lofas, S. SPR screening. Surface plasmon resonance is increasingly useful in the study of biomolecular associations. *Modern Drug Discovery* **2003,** 6, 47-9.

146. Turnbull, W. B.; Daranas, A. H. On the value of c: can low affinity systems be studied by isothermal titration calorimetry? *J Am Chem Soc* **2003,** 125, 14859-66.

147. Tsiodras, S.; Gold, H. S.; Sakoulas, G.; Eliopoulos, G. M.; Wennersten, C.; Venkataraman, L.; Moellering, R. C.; Ferraro, M. J. Linezolid resistance in a clinical isolate of Staphylococcus aureus. *Lancet* **2001,** 358, 207-8.

148. Opal, S. M.; Medeirus, A. A. Molecular mechanisms of antibiotic resistance in bacteria. In *Principles and Practice of Infectious Disease*, 6th ed.; Mandell, G. L.; Bennett, J. C.; Dolin, R., Eds. Elsevier: Philadelphia, **2005**; Vol. 1, pp 253-70.

149. Wright, S. W.; Wrenn, K. D.; Haynes, M. L. Trimethoprim-sulfamethoxazole resistance among urinary coliform isolates. *J Gen Intern Med* **1999,** 14, 606-9.

150. Sievert, D. M.; Rudrik, J. T.; Patel, J. B.; McDonald, L. C.; Wilkins, M. J.; Hageman, J. C. Vancomycin-resistant Staphylococcus aureus in the United States, 2002-2006. *Clin Infect Dis* **2008,** 46, 668-74.

151. Achari, A.; Somers, D. O.; Champness, J. N.; Bryant, P. K.; Rosemond, J.; Stammers, D. K. Crystal structure of the anti-bacterial sulfonamide drug target dihydropteroate synthase. *Nat Struct Biol* **1997,** 4, 490-7.

152. Hampele, I. C.; D'Arcy, A.; Dale, G. E.; Kostrewa, D.; Nielsen, J.; Oefner, C.; Page, M. G.; Schonfeld, H. J.; Stuber, D.; Then, R. L. Structure and function of the dihydropteroate synthase from Staphylococcus aureus. *J Mol Biol* **1997,** 268, 21-30.

153. Baca, A. M.; Sirawaraporn, R.; Turley, S.; Sirawaraporn, W.; Hol, W. G. Crystal structure of Mycobacterium tuberculosis 7,8-dihydropteroate synthase in complex with pterin monophosphate: new insight into the enzymatic mechanism and sulfa-drug action. *J Mol Biol* **2000,** 302, 1193-212.

154. Babaoglu, K.; Qi, J.; Lee, R. E.; White, S. W. Crystal structure of 7,8-dihydropteroate synthase from Bacillus anthracis: mechanism and novel inhibitor design. *Structure* **2004,** 12, 1705-17.

155. Lawrence, M. C.; Iliades, P.; Fernley, R. T.; Berglez, J.; Pilling, P. A.; Macreadie, I. G. The three-dimensional structure of the bifunctional 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase/dihydropteroate synthase of Saccharomyces cerevisiae. *J Mol Biol* **2005,** 348, 655-70.

156. Bagautdinov, B.; Kunishima, N. Crystal structure of dihydropteroate synthase (FolP) from Thermus thermophilus HB8. In RIKEN Structural Genomics/Proteomics Initiative (RSGI): **2006**.

157. Levy, C.; Minnis, D.; Derrick, J. P. Dihydropteroate synthase from Streptococcus pneumoniae: structure, ligand recognition and mechanism of sulfonamide resistance. *Biochem J* **2008**.

158. Domagk, G. Ein beitrag zur chemotherapie der bakteriellen infektionen. *Dtsch Med Wochenschr* **1935,** 61, 250-3.

159. Trefouel, J.; Trefouel, J.; Nitti, F.; Bovet, D. Activite du p-aminophenylsulfamide sur les infections streptococciques experimentales de la souris et du lapin. *C R Soc Biol* **1935,** 120, 756-8.

160. Colebrook, L.; Kenny, M. Treatment of human puerperal infections and of infections in mice with prontosil. *Lancet* **1936,** 1, 1279-86.

161. Woods, D. D. The relation of p-aminobenzoic acid to the mechanism of the action of sulphanilamide. *Br J Exp Pathol* **1940,** 21, 74-90.

162. Fildes, P. A rational approach to research in chemotherapy. *Lancet* **1940,** 235, 955-7.

163. Brown, G. M.; Weisman, R. A.; Molnar, D. A. The biosynthesis of folic acid. *J Biol Chem* **1961,** 236, 2534-43.

164. Brown, G. M. The biosynthesis of folic acid. II. Inhibition by sulfonamides. *J Biol Chem* **1962,** 237, 536-40.

165. Richey, D. P.; Brown, G. M. The biosynthesis of folic acid. IX. Purification and properties of the enzymes required for the formation of dihydropteroic acid. *J Biol Chem* **1969,** 244, 1582-92.

166. Weisman, R. A.; Brown, G. M. The biosynthesis of folic acid. V. Characteristics of the enzyme system that catalyzes the synthesis of dihydropteroic acid. *J Biol Chem* **1964,** 239, 326-31.

167. Bock, L.; Miller, G. H.; Schaper, K. J.; Seydel, J. K. Sulfonamide structure-activity relationships in a cell-free system. 2. Proof for the formation of a sulfonamide-containing folate analog. *J Med Chem* **1974,** 17, 23-8.

168. Roland, S.; Ferone, R.; Harvey, R. J.; Styles, V. L.; Morrison, R. W. The characteristics and significance of sulfonamides as substrates for Escherichia coli dihydropteroate synthase. *J Biol Chem* **1979,** 254, 10337-45.

169. Swedberg, G.; Castensson, S.; Skold, O. Characterization of mutationally altered dihydropteroate synthase and its ability to form a sulfonamide-containing dihydrofolate analog. *J Bacteriol* **1979,** 137, 129-36.

170. Vinnicombe, H. G.; Derrick, J. P. Dihydropteroate synthase from Streptococcus pneumoniae: characterization of substrate binding order and sulfonamide inhibition. *Biochem Biophys Res Commun* **1999,** 258, 752-7.

171. Baca, A. M.; Sirawaraporn, R.; Turley, S.; Sirawaraporn, W.; Hol, W. G. J. Crystal structure of Mycobacterium tuberculosis 6-hydroxymethyl-7,8-dihydropteroate synthase in complex with pterin monophosphate: New insight into the enzymatic mechanism and sulfa-drug action. *Journal of Molecular Biology* **2000,** 302, 1193-1212.

172. Radstrom, P.; Swedberg, G.; Skold, O. Genetic analyses of sulfonamide resistance and its dissemination in gram-negative bacteria illustrate new aspects of R plasmid evolution. *Antimicrob Agents Chemother* **1991,** 35, 1840-8.

173. Perreten, V.; Boerlin, P. A new sulfonamide resistance gene (sul3) in Escherichia coli is widespread in the pig population of Switzerland. *Antimicrob Agents Chemother* **2003,** 47, 1169-72.

174. Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J Comput Chem* **2005,** 26, 1668-88.

175. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* **2003,** 24, 1999-2012.

176. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Antechamber, an accessory software package for molecular mechanics calculations. *J. Mol. Graphics Model.* **2006,** 25, 247-60.

177. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J Comput Chem* **2004,** 25, 1157-74.

178. Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Pairwise solute descreening of solute charges from a dielectric medium. *Chem Phys Lett* **1995,** 246, 122-9.

179. Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J Mol Graph* **1996,** 14, 33-8, 27-8.

180. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **2004,** 25, 1605-12.

181. Giordanetto, F.; Fowler, P. W.; Saqi, M.; Coveney, P. V. Large scale molecular dynamics simulation of native and mutant dihydropteroate synthase-sulphanilamide complexes suggests the molecular basis for dihydropteroate synthase drug resistance. *Philos Transact A Math Phys Eng Sci* **2005,** 363, 2055-73.

182. Onodera, K.; Satou, K.; Hirota, H. Evaluations of molecular docking programs for virtual screening. *J Chem Inf Model* **2007,** 47, 1609-18.

183. Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On evaluating molecular-docking methods for pose prediction and enrichment factors. *J Chem Inf Model* **2006,** 46, 401-15.

184. Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of automated docking programs as virtual screening tools. *J Med Chem* **2005,** 48, 962-76.

185. Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J Med Chem* **2006,** 49, 5912-31.

186. Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J Med Chem* **2004,** 47, 45-55.

187. Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: comparative data on docking algorithms. *J Med Chem* **2004,** 47, 558-65.

188. Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. Evaluation of library ranking efficacy in virtual screening. *J Comput Chem* **2005,** 26, 11-22.

189. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004,** 57, 225-42.

190. Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004,** 56, 235-49.

191. Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. Evaluation and application of multiple scoring functions for a virtual screening experiment. *J Comput Aided Mol Des* **2004,** 18, 333-44.

192. Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J Mol Model* **2003,** 9, 47-57.

193. Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L., 3rd. Comparative study of several algorithms for flexible ligand docking. *J Comput Aided Mol Des* **2003,** 17, 755-63.

194. Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem* **2003,** 46, 2287-303.

195. Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J Med Chem* **2001,** 44, 1035-42.

196. Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem* **2000,** 43, 4759-67.

197. Cole, J. C.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins* **2005,** 60, 325-32.

198. Jain, A. N. Bias, reporting, and sharing: computational evaluations of docking methods. *J Comput Aided Mol Des* **2008,** 22, 201-12.

199. Neyman, J.; Pearson, E. S. On the problem of the most efficient tests of statistical hypotheses. *Philos Trans R Soc, London, Ser A* **1933,** 231, 289-337.

200. Neyman, J.; Pearson, E. S. The testing of statistical hypotheses in relation to probabilities a priori. *Proc. Cambridge Philos Soc* **1933,** 20, 492-510.

201. Swets, J. A.; Dawes, R. M.; Monahan, J. Better decisions through science. *Sci Am* **2000,** 283, 82-7.

202. Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J Med Chem* **2005,** 48, 2534-47.

203. Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J Chem Inf Model* **2007,** 47, 488-508.

204. Morrison, R. W.; Styles, V. L. Pyrimido[4,5-c]pyridazines. 5. Summary of cyclizations with vicinally functionalized reagents and studies of the reductive behavior of the ring system. *J Org Chem* **1982,** 47, 674-80.

205. Mallory, W. R.; Morrison, R. W.; Styles, V. L. Pyrimido[4,5-c]pyridazines. 3. Preferential formation of 8-amino-1H-pyrimido[4,5-c]-1,2-diazepin-6(7H)-ones by cyclizations with alpha,gamma-diketo Esters. *J Org Chem* **1982,** 47, 667-74.

206. Lever, O. W., Jr.; Bell, L. N.; McGuire, H. M.; Ferone, R. Monocyclic pteridine analogues. Inhibition of Escherichia coli dihydropteroate synthase by 6-amino-5-nitrosoisocytosines. *J Med Chem* **1985,** 28, 1870-4.

207. Lever, O. W., Jr.; Bell, L. N.; Hyman, C.; McGuire, H. M.; Ferone, R. Inhibitors of dihydropteroate synthase: substituent effects in the side-chain aromatic ring of 6-[[3-(aryloxy)propyl]amino]-5-nitrosoisocytosines and synthesis and inhibitory potency of bridged 5-nitrosoisocytosine-p-aminobenzoic acid analogues. *J Med Chem* **1986,** 29, 665-70.

208. Sybyl 7.9, Tripos International, 1699 South Hanley Rd., St. Louis, Missiouri, 63144, USA

209. GOLD v3.1.1, Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK

210. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem* **2004,** 47, 1750-9.

211. Glide 4.0, Schrodinger, Inc.l, 120 West 45th St., 29th Floor, New York, NY 10036-4041, USA

212. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J Mol Biol* **1982,** 161, 269-88.

213. Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and validation of a modular, extensible docking program: DOCK 5. *J Comput Aided Mol Des* **2006,** 20, 601-19.

214. Binder, K. The Monte Carlo method in condensed matter physics. In *Topics in Applied Physics*, Springer: Berlin, **1993**; Vol. 71, pp 1-392.

215. Pearlman, R. S. Rapid Generation of high quality approximate 3-dimensional molecular structures. *Chem Des Auto News* **1987,** 2, 1-7.

216. Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. Conjugate gradients. In *Numerical Recipes in C, The Art of Scientific Computing*, Cambridge University Press: Cambridge, **1988**; p 312.

217. Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein-ligand interactions using negative training data. *J Med Chem* **2006,** 49, 5856-68.

218. Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J Chem Inf Comput Sci* **2004,** 44, 793-806.

219. Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J Chem Inf Comput Sci* **2001,** 41, 1395-406.

220. Irwin, J. J.; Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **2005,** 45, 177-82.

221. Powell, M. J. D. Restart Procedures for the conjugate gradient method. *Mathematical Programming* **1977,** 12, 241-54.

222. Martin, Y. C. 3D database searching in drug design. *J Med Chem* **1992,** 35, 2145-54.

223. Hurst, T. Flexible 3D Searching: The directed tweak technique. *J Chem Inf Comput Sci* **1994,** 34, 190-6.

224. R. S. Pearlman, "Concord", distributed by Tripos International, St. Louis, Missouri, 43144, USA.

225. Frothingham, R.; Stout, J. E.; Hamilton, C. D. Current issues in global tuberculosis control. *Int J Infect Dis* **2005,** 9, 297-311.

226. WHO. *Global Tuberculosis Control: Surveillance, Planning, Financing: WHO Report 2007*; World Health Organization: Geneva, **2007**.

227. Ginsberg, A. M.; Spigelman, M. Challenges in tuberculosis drug research and development. *Nature Medicine* **2007,** 13, 290-4.

228. Sacchettini, J. C.; Rubin, E. J.; Freundlich, J. S. Drugs versus bugs: in pursuit of the persistent predator Mycobacterium tuberculosis. *Nat Rev Microbiol* **2008,** 6, 41-52.

229. Tangallapally, R. P.; Yendapally, R.; Lee, R. E.; Hevener, K.; Jones, V. C.; Lenaerts, A. J.; McNeil, M. R.; Wang, Y.; Franzblau, S.; Lee, R. E. Synthesis and evaluation of nitrofuranylamides as novel antituberculosis agents. *J Med Chem* **2004,** 47, 5276-83.

230. Tangallapally, R. P.; Yendapally, R.; Lee, R. E.; Lenaerts, A. J. Synthesis and evaluation of cyclic secondary amine substituted phenyl and benzyl nitrofuranyl amides as novel antituberculosis agents. *J Med Chem* **2005,** 48, 8261-9.

231. Tangallapally, R. P.; Lee, R. E.; Lenaerts, A. J.; Lee, R. E. Synthesis of new and potent analogues of anti-tuberculosis agent 5-nitro-furan-2-carboxylic acid 4-(4-benzyl-piperazin-1-yl)-benzylamide with improved bioavailability. *Bioorg Med Chem Lett* **2006,** 16, 2584-9.

232. Tangallapally, R. P.; Yendapally, R.; Daniels, A. J.; Lee, R. E.; Lee, R. E. Nitrofurans as novel anti-tuberculosis agents: identification, development and evaluation. *Curr Top Med Chem* **2007,** 7, 509-26.

233. Budha, N. R.; Mehrotra, N.; Tangallapally, R.; Rakesh; Daniels, A.; Lee, R. E.; Meibohm, B. Pharmacokinetically-guided lead optimization for nitrofuranylamides against tuberculosis. In *The AAPS Journal (in Press)*, **2008**.

234. Hurdle, J. G.; Lee, R. B.; Budha, N. R.; Carson, E. I.; Qi, J.; McNeil, M. R.; Lenaerts, A. J.; Franzblau, S. G.; Meibohm, B.; Lee, R. E. A microbiological assessment of novel nitrofuranylamides as anti-tuberculosis agents In *Antimicrob Agents Chemother (In Press)*, **2008**.

235. Budha, N. R. L., R. E.; Meibohm, B. . Biopharmaceutics, pharmacokinetics and pharmacodynamics of antituberculosis drugs. *Current Medicinal Chemistry (in press)* **2008**.

236. Manjunatha, U. H.; Boshoff, H.; Dowd, C. S.; Zhang, L.; Albert, T. J.; Norton, J. E.; Daniels, L.; Dick, T.; Pang, S. S.; Barry, C. E., 3rd. Identification of a nitroimidazo-oxazine-specific protein involved in PA-824 resistance in Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* **2006,** 103, 431-6.

237. Matsumoto, M.; Hashizume, H.; Tomishige, T.; Kawasaki, M.; Tsubouchi, H.; Sasaki, H.; Shimokawa, Y.; Komatsu, M. OPC-67683, a nitro-dihydro-imidazooxazole derivative with promising action against tuberculosis in vitro and in mice. *PLoS Med* **2006,** 3, e466.

238. Barry, C. E., 3rd; Boshoff, H. I.; Dowd, C. S. Prospects for clinical introduction of nitroimidazole antibiotics for the treatment of tuberculosis. *Curr Pharm Des* **2004,** 10, 3239-62.

239. Ventura, C.; Martins, F. Application of quantitative structure-activity relationships to the modeling of antitubercular compounds. 1. The hydrazide family. *J Med Chem* **2008,** 51, 612-24.

240. Gupta, R. A.; Gupta, A. K.; Soni, L. K.; Kaskhedikar, S. G. Rationalization of physicochemical characters of oxazolyl thiosemicarbazone analogs towards multi-drug resistant tuberculosis: a QSAR approach. *Eur J Med Chem* **2007,** 42, 1109-16.

241. Saquib, M.; Gupta, M. K.; Sagar, R.; Prabhakar, Y. S.; Shaw, A. K.; Kumar, R.; Maulik, P. R.; Gaikwad, A. N.; Sinha, S.; Srivastava, A. K.; Chaturvedi, V.; Srivastava, R.; Srivastava, B. S. C-3 alkyl/arylalkyl-2,3-dideoxy hex-2-enopyranosides as antitubercular agents: synthesis, biological evaluation, and QSAR study. *J Med Chem* **2007,** 50, 2942-50.

242. Nayyar, A.; Monga, V.; Malde, A.; Coutinho, E.; Jain, R. Synthesis, anti-tuberculosis activity, and 3D-QSAR study of 4-(adamantan-1-yl)-2-substituted quinolines. *Bioorg Med Chem* **2007,** 15, 626-40.

243. Cramer, R. D., 3rd; Patterson, D. E.; Bunce, J. D. Recent advances in comparative molecular field analysis (CoMFA). *Prog Clin Biol Res* **1989,** 291, 161-5.

244. Thibaut, U.; Folkers, G.; Klebe, G.; Kubinyi, H.; Merz, A.; Rognan, D. Recommendations for CoMFA studies and 3D QSAR publications. *Quant Struct-Act Relat* **1994,** 13, 1-3.

245. Sybyl. *Sybyl 8.0*, 8.0; Tripos, Inc.: St. Louis, MO, **2007**.

246. ChemBioOffice. *ChemBioOffice Ultra 2008*, 11; CambridgeSoft: **2008**.

247. Marvin. *Marvin 4.1.13*, 4.1.13; ChemAxon: **2007**.

248. Stewart, J. J. P. Optimization of parameters for semi-empirical methods I-Method. *J Comp Chem* **1989,** 10.

249. Li, X.; Manjunatha, U. H.; Goodwin, M. B.; Knox, J. E.; Lipinski, C. A.; Keller, T. H.; Barry, C. E., 3rd; Dowd, C. S. Synthesis and antitubercular activity of 7-(R)- and 7-(S)-methyl-2-nitro-6-(S)-(4-(trifluoromethoxy)benzyloxy)-6,7-dihydro-5H-imidazo[2,1-b][1,3]oxazines, analogues of PA-824. *Bioorg Med Chem Lett* **2008,** 18, 2256-62.

250. Bush, B. L.; Nachbar, R. B., Jr. Sample-distance partial least squares: PLS optimized for many variables, with application to CoMFA. *J Comput Aided Mol Des* **1993,** 7, 587-619.

251. Clark, R. D.; Fox, P. C. Statistical variation in progressive scrambling. *J Comput Aided Mol Des* **2004,** 18, 563-76.

252. Clark, R. D.; Sprous, D. G.; Leonard, J. M. Validating models based on large data sets. In *Rational Approaches to Drug Design*, Holtje, H.-D.; Sippl, W., Eds. Prous Science SA: **2001**; pp 475-85.

253. Luco, J. M.; Ferretti, F. H. QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives. *J Chem Inf Comput Sci* **1997,** 37, 392-401.

254. Datar, P.; Desai, P.; Coutinho, E.; Iyer, K. CoMFA and CoMSIA studies of angiotensin (AT1) receptor antagonists. *J Mol Model.* **2002,** 10, 290-301.

255. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSAR *Environ Health Perspect* **2003,** 111, 1361-75.

256. Holliday, J. D.; Ranade, S. S.; WIllett, P. A fast algorithm for selecting sets of dissimilar molecules from large chemical databases. *Quant Struct Act Relat* **1996,** 14, 501-6.

257. Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenter*. Wiley: New York, **1978**.

258. Davis, A. M.; Teague, S. J.; Kleywegt, G. J. Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew Chem Int Ed Engl* **2003,** 42, 2718-36.

259. Yu, H.; Adedoyin, A. ADME-Tox in drug discovery: integration of experimental and computational technologies. *Drug Discov Today* **2003,** 8, 852-61.

260. Johnson, S. R. The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J Chem Inf Model* **2008,** 48, 25-6.

261. Boczko, E. M.; Brooks, C. L., 3rd. First-principles calculation of the folding free energy of a three-helix bundle protein. *Science* **1995,** 269, 393-6.

262. Hansmann, U. H. E. Parallel tempering algorithm for conformational studies of biological molecules. *Chem Phys Lett* **1997,** 281, 140-50.

263. Faller, R.; Yan, Q. L.; de Pablo, J. J. Multicanonical parallel tempering. *J Chem Phys* **2002,** 116, 249-53.

264. Rich, D. H. Pepstatin-derived inhibitors of aspartic proteinases. A close look at an apparent transition-state analogue inhibitor. *J Med Chem* **1985,** 28, 263-73.

# APPENDICES

## A. Molecular Dynamics Force Field Parameter Files for Non-Standard Residues

### A.1 Pterin-SMX Parameter/Topology File

```
    0    0    2

This is a remark line
molecule.res
N5_     INT  0
CORRECT     OMIT DU   BEG
  0.0000
   1  DUMM  DU    M    0  -1  -2     0.000      .0        .0      .00000
   2  DUMM  DU    M    1   0  -1     1.449      .0        .0      .00000
   3  DUMM  DU    M    2   1   0     1.522   111.1        .0      .00000
   4  C13   ca    M    3   2   1     1.540   111.208   180.000  -0.00433
   5  H8    ha    E    4   3   2     1.082    83.102   156.178   0.14046
   6  C14   ca    M    4   3   2     1.399   105.293   -84.387  -0.21466
   7  H7    ha    E    6   4   3     1.085   119.228   -91.706   0.11383
   8  C9    ca    M    6   4   3     1.401   120.514    88.195   0.12742
   9  N5    nh    B    8   6   4     1.358   120.417   178.992  -0.80927
  10  H10   hn    E    9   8   6     1.030   120.607    -0.996   0.36522
  11  H9    hn    E    9   8   6     1.031   120.529   179.754   0.36546
  12  C10   ca    M    8   6   4     1.400   119.054    -1.033  -0.21394
  13  H6    ha    E   12   8   6     1.085   120.403  -177.889   0.11437
  14  C11   ca    M   12   8   6     1.398   120.440     1.908  -0.00262
  15  H5    ha    E   14  12   8     1.084   120.606   179.160   0.14670
  16  C12   ca    M   14  12   8     1.397   120.232    -0.948  -0.31054
  17  S15   sy    M   16  14  12     1.772   119.270  -178.843   1.45835
  18  O2    o     E   17  16  14     1.503   110.057    84.879  -0.69067
  19  O3    o     E   17  16  14     1.494   110.394   -38.078  -0.69703
  20  N7    ne    M   17  16  14     1.701   104.212  -156.868  -0.98682
  21  C18   cc    M   20  17  16     1.357   122.568    80.430   0.59964
  22  C19   cc    S   21  20  17     1.459   121.516   146.895  -0.36842
  23  H1    ha    E   22  21  20     1.090   129.891    -0.290   0.15540
  24  N10   nd    M   21  20  17     1.277   124.888   -33.961  -0.30744
  25  O5    os    M   24  21  20     1.415   105.738  -179.774  -0.17914
  26  C20   cd    M   25  24  21     1.327   105.707     0.559   0.19674
  27  C21   c3    M   26  25  24     1.504   117.908   179.654  -0.11856
  28  H4    hc    E   27  26  25     1.099   109.347    77.381   0.03945
  29  H3    hc    E   27  26  25     1.100   110.775  -162.295   0.03653
  30  H2    hc    E   27  26  25     1.100   109.848   -42.156   0.04387


LOOP
  C12  C13
  C20  C19

IMPROPER
  C12  C14  C13   H8
  C13   C9  C14   H7
```

```
  C10  C14   C9   N5
   C9   H9   N5  H10
  C11   C9  C10   H6
  C10  C12  C11   H5
  C11  C13  C12  S15
  C19  N10  C18   N7
  C18  C20  C19   H1
  C21  C19  C20   O5

DONE
STOP
```

### A.2 Pterin-SMX Additional Parameters

```
remark goes here
MASS

BOND
ne-cc  381.80   1.414       same as ce-ne

ANGLE
ca-sy-ne   40.139      98.980   Calculated with empirical approach
sy-ne-cc   62.900     114.810   same as c2-ne-sy
ne-cc-cc   69.300     121.150   same as cc-cc-n2
ne-cc-nd   78.000     113.820   same as n2-c2-n2

DIHE
sy-ne-cc-cc   1    0.800       180.000           2.000      same as X -ce-
ne-X
sy-ne-cc-nd   1    0.800       180.000           2.000      same as X -ce-
ne-X
cc-nd-os-cd   1    3.000       180.000           2.000      same as X -ne-
os-X
cc-cd-os-nd   1    1.050       180.000           2.000      same as X -c2-
os-X
nd-os-cd-c3   1    1.050       180.000           2.000      same as X -c2-
os-X

IMPROPER
ca-ca-ca-ha         1.1          180.0         2.0          General
improper torsional angle (2 general atom types)
ca-ca-ca-nh         1.1          180.0         2.0          Using default
value
ca-hn-nh-hn         1.1          180.0         2.0          Using default
value
ca-ca-ca-sy         1.1          180.0         2.0          Using default
value
cc-nd-cc-ne         1.1          180.0         2.0          Using default
value
cc-cd-cc-ha         1.1          180.0         2.0          Using default
value
c3-cc-cd-os         1.1          180.0         2.0          Using default
value
```

```
NONBON
```

## A.3 *p*ABA Parameter/Topology File

```
    0    0    2

This is a remark line
molecule.res
<1>    INT  0
CORRECT     OMIT DU   BEG
  0.0000
   1  DUMM  DU    M    0  -1  -2     0.000      .0        .0      .00000
   2  DUMM  DU    M    1   0  -1     1.449      .0        .0      .00000
   3  DUMM  DU    M    2   1   0     1.522   111.1        .0      .00000
   4  O23   o     M    3   2   1     1.540   111.208   180.000  -0.83558
   5  C21   c     M    4   3   2     1.262    74.977    45.765   0.90932
   6  O22   o     E    5   4   3     1.261   121.626    27.993  -0.83588
   7  C20   ca    M    5   4   3     1.515   119.163  -152.033  -0.14721
   8  C19   ca    M    7   5   4     1.399   120.483    -0.007  -0.08417
   9  H13   ha    E    8   7   5     1.086   119.720     0.158   0.14961
  10  C17   ca    M    8   7   5     1.398   120.495  -179.805  -0.20197
  11  H11   ha    E   10   8   7     1.085   119.168   179.918   0.10509
  12  C15   ca    M   10   8   7     1.400   120.554    -0.037   0.06798
  13  N14   nh    B   12  10   8     1.356   120.602   179.891  -0.79670
  14  H9    hn    E   13  12  10     1.031   120.600     0.120   0.35081
  15  H8    hn    E   13  12  10     1.031   120.658  -179.768   0.35081
  16  C16   ca    M   12  10   8     1.401   118.888    -0.172  -0.20256
  17  H10   ha    E   16  12  10     1.085   120.283  -179.870   0.10494
  18  C18   ca    M   16  12  10     1.398   120.504     0.160  -0.08410
  19  H12   ha    E   18  16  12     1.085   119.770  -179.979   0.14961


LOOP
  C18  C20

IMPROPER
  C20  O23  C21  O22
  C21  C18  C20  C19
  C20  C17  C19  H13
  C15  C19  C17  H11
  C16  C17  C15  N14
  C15   H9  N14   H8
  C15  C18  C16  H10
  C16  C20  C18  H12

DONE
STOP
```

## A.4 *p*ABA Additional Parameters

```
remark goes here
MASS

BOND

ANGLE

DIHE

IMPROPER
ca-o -c -o          1.1          180.0         2.0          General
improper torsional angle (1 general atom type)
c -ca-ca-ca         1.1          180.0         2.0          Using default
value
ca-ca-ca-ha         1.1          180.0         2.0          General
improper torsional angle (2 general atom types)
ca-ca-ca-nh         1.1          180.0         2.0          Using default
value
ca-hn-nh-hn         1.1          180.0         2.0          Using default
value

NONBON
```

## A.5 DHPP Parameter/Topology File

```
    0    0    2

This is a remark line
molecule.res
PT1    INT  0
CORRECT     OMIT DU   BEG
  0.0000
   1  DUMM  DU    M    0  -1  -2     0.000      .0        .0      .00000
   2  DUMM  DU    M    1   0  -1     1.449      .0        .0      .00000
   3  DUMM  DU    M    2   1   0     1.522   111.1        .0      .00000
   4  N11   nh    M    3   2   1     1.540   111.208   180.000  -0.86704
   5  H1    hn    E    4   3   2     1.031   150.274   -32.291   0.36913
   6  H2    hn    E    4   3   2     1.029    68.154    78.161   0.35130
   7  C7    cd    M    4   3   2     1.305    61.935  -132.431   0.56438
   8  N9    nc    E    7   4   3     1.272   119.636   146.126  -0.69312
   9  N4    n     M    7   4   3     1.308   119.085   -33.959  -0.49023
  10  H5    hn    E    9   7   4     1.027   120.880    -0.133   0.30570
  11  C2    c     M    9   7   4     1.305   121.610   179.827   0.71690
  12  O1    o     E   11   9   7     1.220   119.655   179.814  -0.68699
  13  C3    cd    M   11   9   7     1.477   120.189    -0.283  -0.06703
  14  N6    nf    E   13  11   9     1.446   122.983  -178.373  -0.62161
  15  C5    cc    M   13  11   9     1.336   117.276     0.945   0.33507
  16  N8    nh    M   15  13  11     1.304   121.701   179.156  -0.67478
  17  H6    hn    E   16  15  13     1.032   118.656   178.581   0.41204
  18  C12   c3    M   16  15  13     1.466   122.289    -2.096  -0.04144
```

```
  19  H3    h1    E   18  16  15     1.102   110.446   129.399   0.20505
  20  H4    h1    E   18  16  15     1.099   106.410  -111.398   0.18287
  21  C10   c2    M   18  16  15     1.505   113.974     5.741   0.47902
  22  C13   c3    M   21  18  16     1.508   119.414   173.133   0.14476
  23  H7    h1    E   22  21  18     1.096   108.569   -90.894   0.13155
  24  H14   h1    E   22  21  18     1.098   106.819   153.769  -0.00181
  25  O38   os    M   22  21  18     1.437   116.106    36.793  -0.58527
  26  P40   p5    M   25  22  21     1.606   135.494   -73.898   1.56950
  27  O2    o     E   26  25  22     1.491   112.442    68.982  -0.91530
  28  O3    o     E   26  25  22     1.492   104.553  -179.581  -0.90410
  29  O4    os    M   26  25  22     1.606   116.062   -61.355  -0.82164
  30  P44   p5    M   29  26  25     1.605   119.125    36.344   1.46331
  31  O6    o     E   30  29  26     1.491   108.993   165.384  -0.95009
  32  O7    o     E   30  29  26     1.490   113.198    45.487  -0.96046
  33  O5    o     M   30  29  26     1.491   110.520   -77.439  -0.94965


LOOP
   C5   N9
  C10   N6

IMPROPER
   C7   H1  N11   H2
   N4   N9   C7  N11
   C2   C7   N4   H5
   C3   N4   C2   O1
   C2   C5   C3   N6
   C3   N9   C5   N8
  C12   C5   N8   H6
  C12  C13  C10   N6

DONE
STOP
```

### A.6 DHPP Additional Parameters

```
remark goes here
MASS

BOND
cd-nf  597.70   1.280       same as c2-nf

ANGLE
c -cd-nf   68.400     120.890   same as c -cd-n2
cd-nf-c2   70.800     118.180   same as c2-n2-c2
nf-cd-cc   71.300     126.010   same as c2-c2-n2
nh-c3-c2   67.571     107.790   Calculated with empirical approach

DIHE
c -cd-nf-c2   1    4.150       180.000           2.000      same as X -c2-
nf-X
cc-cd-nf-c2   1    4.150       180.000           2.000      same as X -c2-
nf-X
```

```
IMPROPER
cd-hn-nh-hn         1.1          180.0         2.0          Using default
value
n -nc-cd-nh         1.1          180.0         2.0          Using default
value
c -cd-n -hn         1.1          180.0         2.0          General
improper torsional angle (2 general atom types)
cd-n -c -o         10.5          180.0         2.0          General
improper torsional angle (2 general atom types)
c -cc-cd-nf         1.1          180.0         2.0          Using default
value
cd-nc-cc-nh         1.1          180.0         2.0          Using default
value
c3-cc-nh-hn         1.1          180.0         2.0          Using default
value
c3-c3-c2-nf         1.1          180.0         2.0          Using default
value

NONBON
```

### A.7 PPi Parameter/Topology File

```
    0    0    2

This is a remark line
molecule.res
PPI    INT  0
CORRECT     OMIT DU   BEG
  0.0000
   1  DUMM  DU    M    0  -1  -2     0.000      .0        .0      .00000
   2  DUMM  DU    M    1   0  -1     1.449      .0        .0      .00000
   3  DUMM  DU    M    2   1   0     1.522   111.1        .0      .00000
   4  O1    o     M    3   2   1     1.540   111.208   180.000  -1.02198
   5  P3    p5    M    4   3   2     1.489   128.443   117.576   1.49384
   6  O10   o     E    5   4   3     1.485   111.856  -155.385  -1.02198
   7  O11   o     E    5   4   3     1.487   111.295   -26.083  -1.02198
   8  O3    os    M    5   4   3     1.647   101.563    87.357  -0.85581
   9  PO4   p5    M    8   5   4     1.647   131.548   161.142   1.49384
  10  O5    o     E    9   8   5     1.485   109.960    42.494  -1.02198
  11  O6    o     E    9   8   5     1.487   106.882   -82.185  -1.02198
  12  O4    o     M    9   8   5     1.489   101.559   161.071  -1.02198


LOOP

IMPROPER

DONE
STOP
```

## A.8 $SO_4$ Parameter/Topology File

```
    0    0    2

This is a remark line
molecule.res
SO4     INT  0
CORRECT     OMIT DU   BEG
  0.0000
   1  DUMM  DU    M    0  -1  -2     0.000      .0        .0      .00000
   2  DUMM  DU    M    1   0  -1     1.449      .0        .0      .00000
   3  DUMM  DU    M    2   1   0     1.522   111.1        .0      .00000
   4  O4    o     M    3   2   1     1.540   111.208   180.000  -0.88594
   5  S     s6    M    4   3   2     1.420    97.201   111.356   1.54374
   6  O2    o     E    5   4   3     1.423   119.786   -82.939  -0.88593
   7  O1    o     E    5   4   3     1.410   118.162    43.175  -0.88596
   8  O3    o     M    5   4   3     1.399   117.454   154.257  -0.88592


LOOP

IMPROPER

DONE
STOP
```

# B. Chapter 2 Supplemental Material

## B.1 Molecular Dynamics Equilibrium Energy Plots

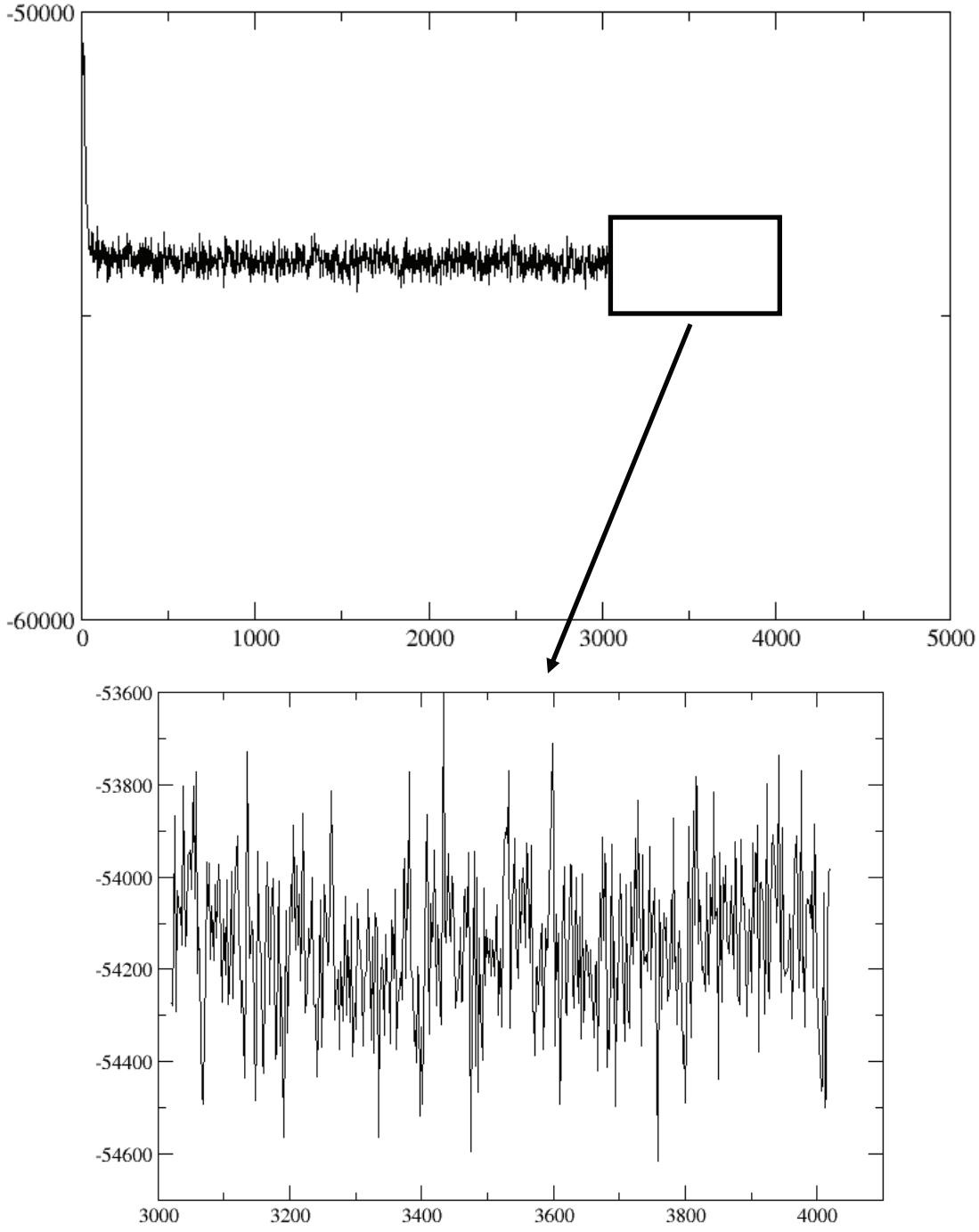



**Figure B.1. DHPS Pterin-SMX 4ns, Explicit Simulation Total Energy Plot**

## B.2 DHPS Molecular Dynamics RMSD Calculations

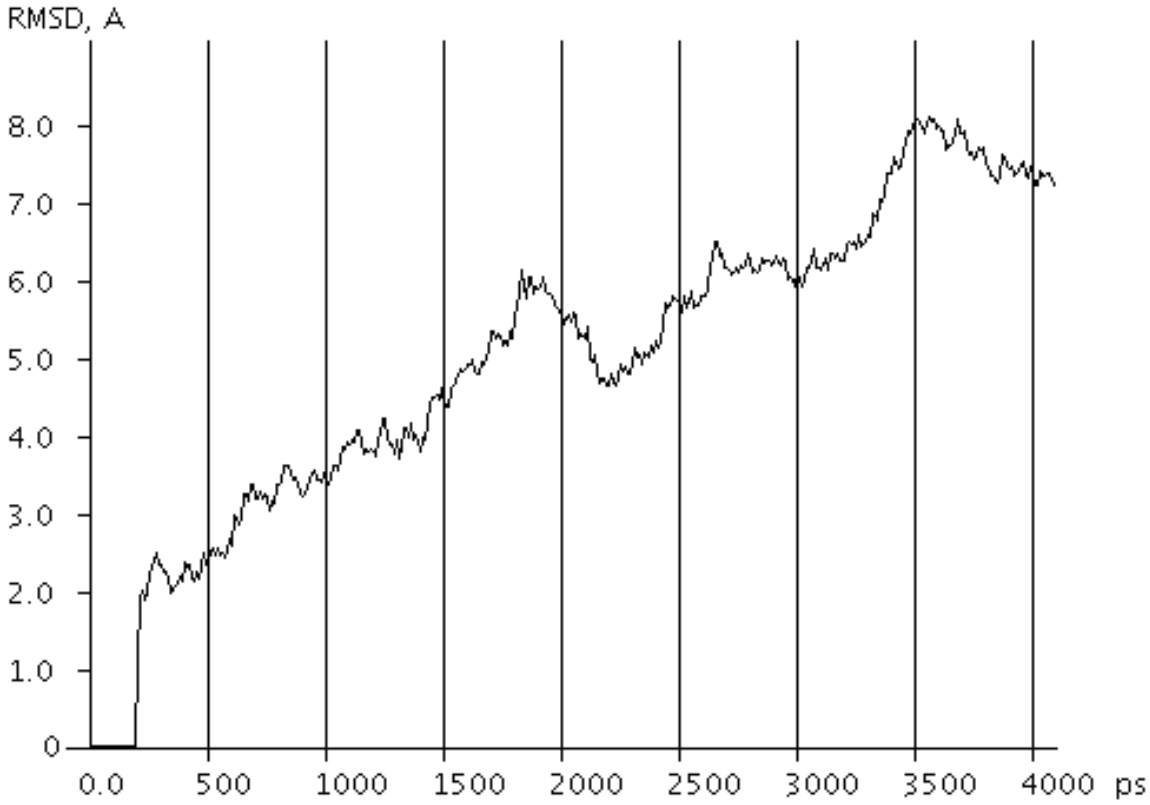



**Figure B.2. DHPS RMSD Calculation, Full Enzyme**

This is a standard RMSD calculation plot referenced to initial frame DHPS simulation 2-17. In this figure, RMSD is calculated for backbone atoms of the entire protein.


**Figure B.3. DHPS RMSD Calculation, Loop 1**

This is a standard RMSD calculation plot referenced to initial frame DHPS simulation 2-17. In this figure, RMSD is calculated for all atoms of loop 1, from Val28 to Val40.

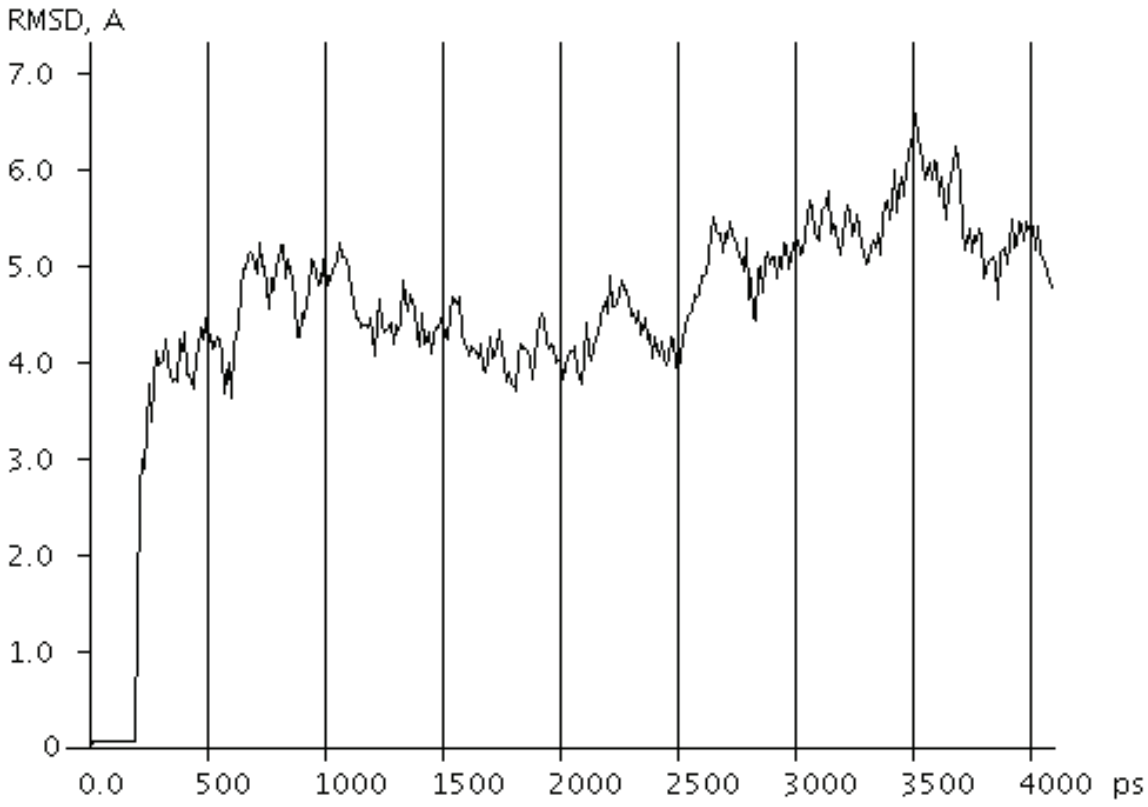



**Figure B.4. DHPS RMSD Calculation, Loop 2**

This is a standard RMSD calculation plot referenced to initial frame DHPS simulation 2-17. In this figure, RMSD is calculated for all atoms of loop 2, from Glu65 to GLU77.

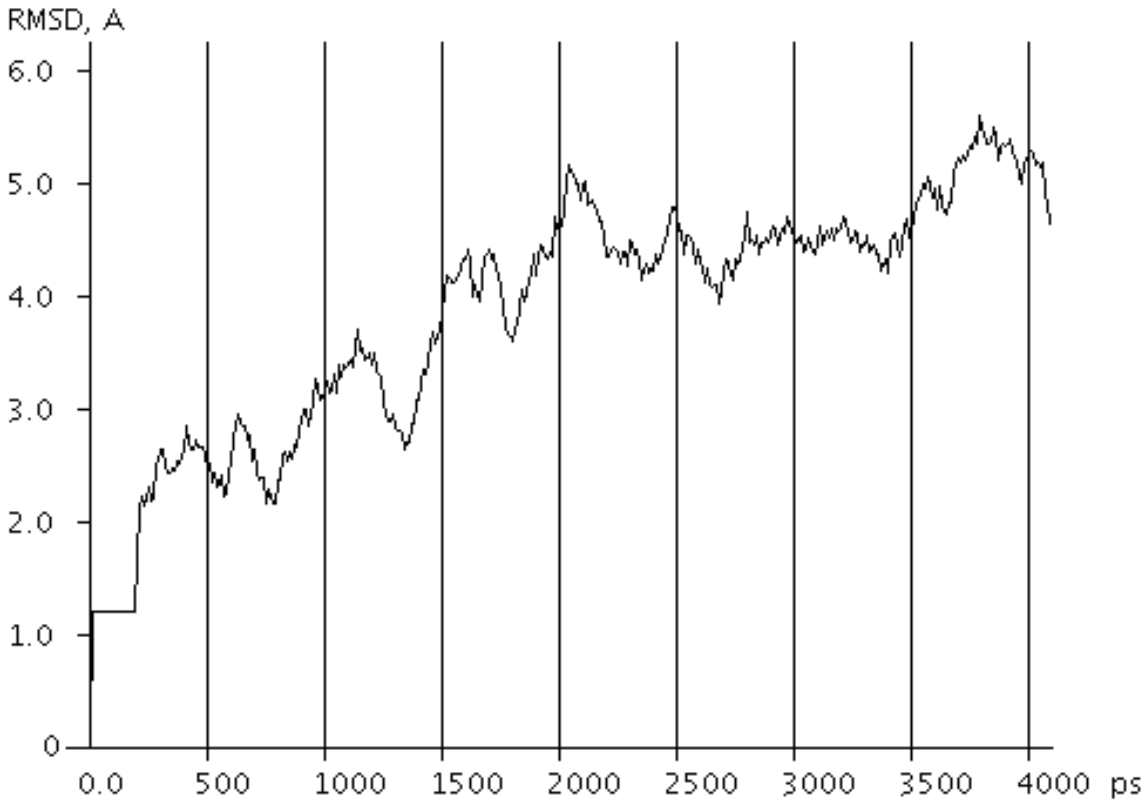



**Figure B.5. DHPS RMSD Calculation, Helices and β Strands**

This is a standard RMSD calculation plot referenced to initial frame DHPS simulation 2-17. In this figure, RMSD is calculated for all backbone atoms of the helix and strand secondary structures and excludes all atoms in flexible loop regions.

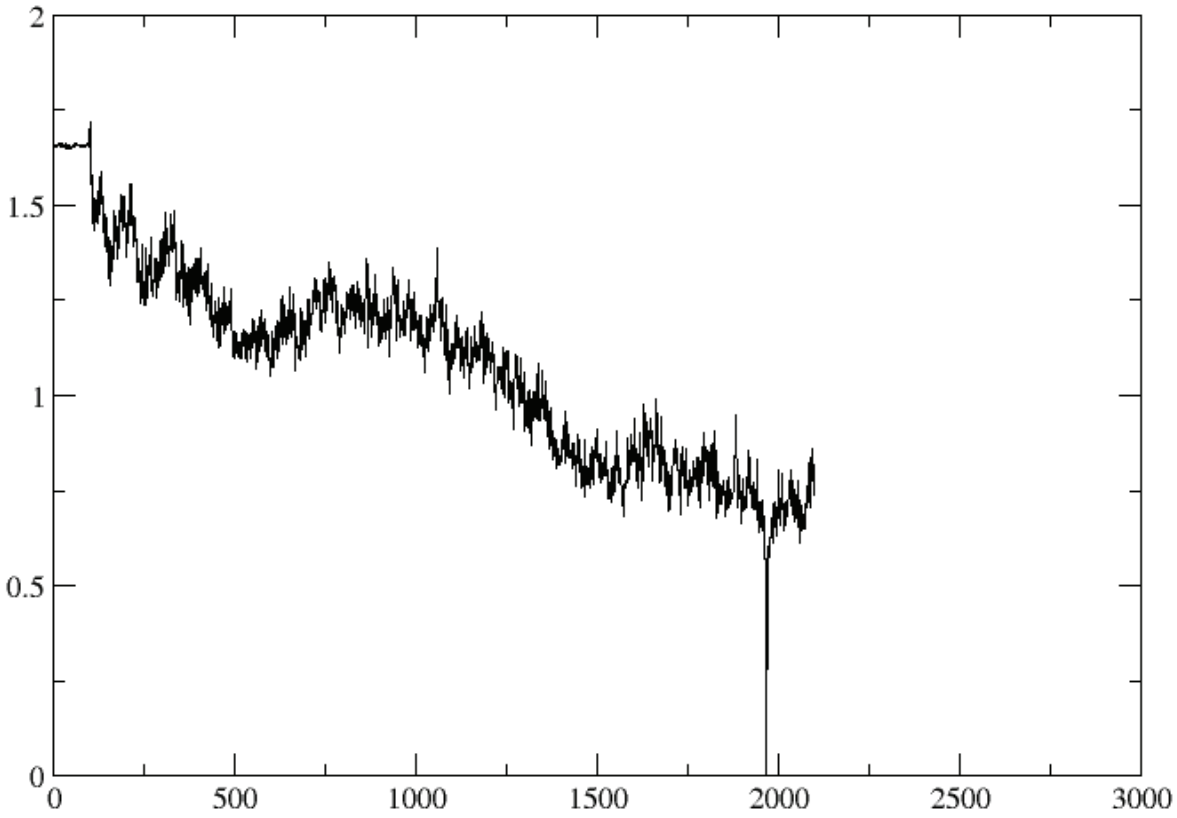



**Figure B.6. DHPS RMSD Calculation, Full Protein to Minimum Energy Structure**

This is a standard RMSD calculation plot referenced to the lowest energy conformation from DHPS simulation 2-17. In this figure, RMSD is calculated for backbone atoms of the entire protein.

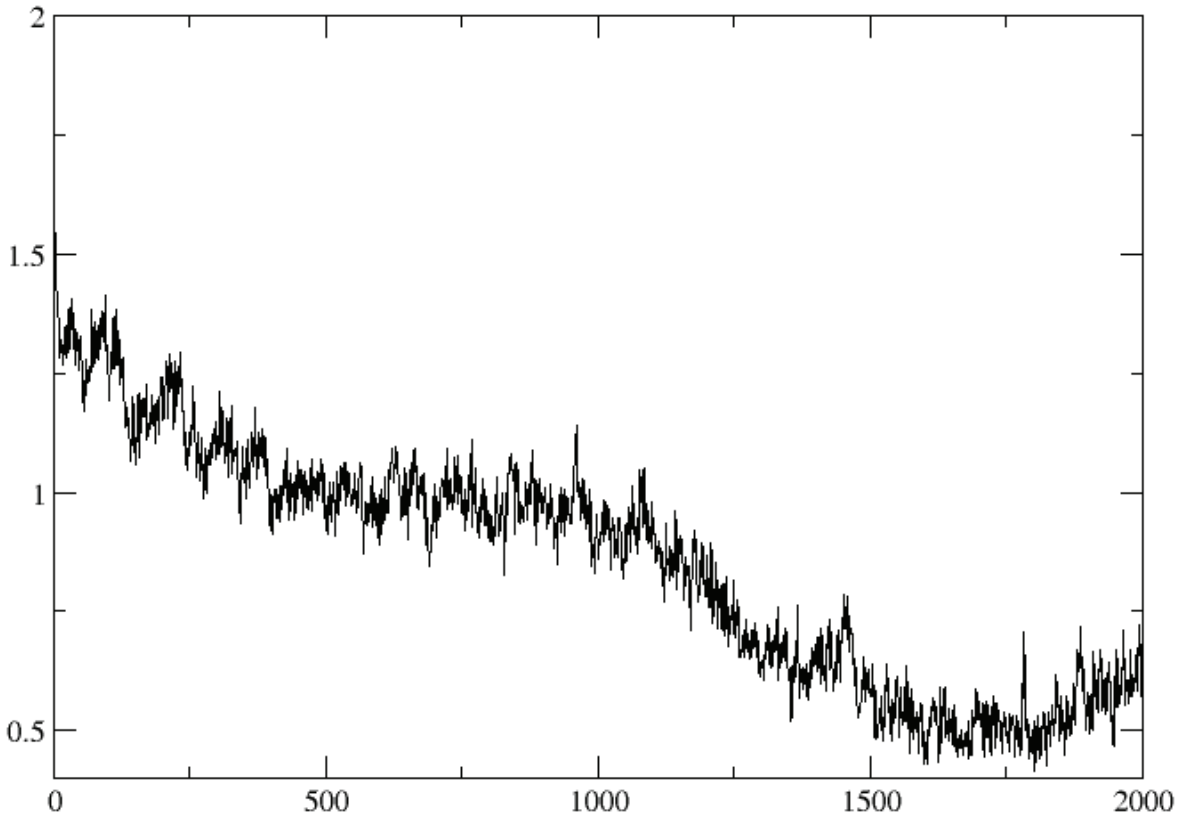



**Figure B.7. DHPS RMSD Calculation, Full Protein to Average Structure**

This is a standard RMSD calculation plot referenced to the average conformation from DHPS simulation 2-17. In this figure, RMSD is calculated for backbone atoms of the entire protein.

**Table B.1. DHPS Average and Minimum Energy Structure RMSD Values**

| DHPS MD Structure | RMSD Calculation | RMSD (Å) Value |
|---|---|---|
| Minimum Energy Structure | All Backbone Atoms | 3.091 |
| | Loop 2 Backbone Atoms | 3.017 |
| | Loop 2 All Atoms | 4.321 |
| Average Structure | All Backbone Atoms | 2.600 |
| | Loop 2 Backbone Atoms | 3.055 |
| | Loop 2 All Atoms | 4.388 |

The RMSD values reported above are referenced to a DHPS crystal structure that was solved by our group after our MD studies were completed. In this new structure, loop 2 is visible in its entirety, although there is no ligand bound in the pterin binding site. Loop 1 is missing from the new structure, indicating that the crystal contact has been lost and the loop is highly mobile.

## B.3 Trajectory Analysis

### *B.3.1 Phe33 Dihedral Analysis*

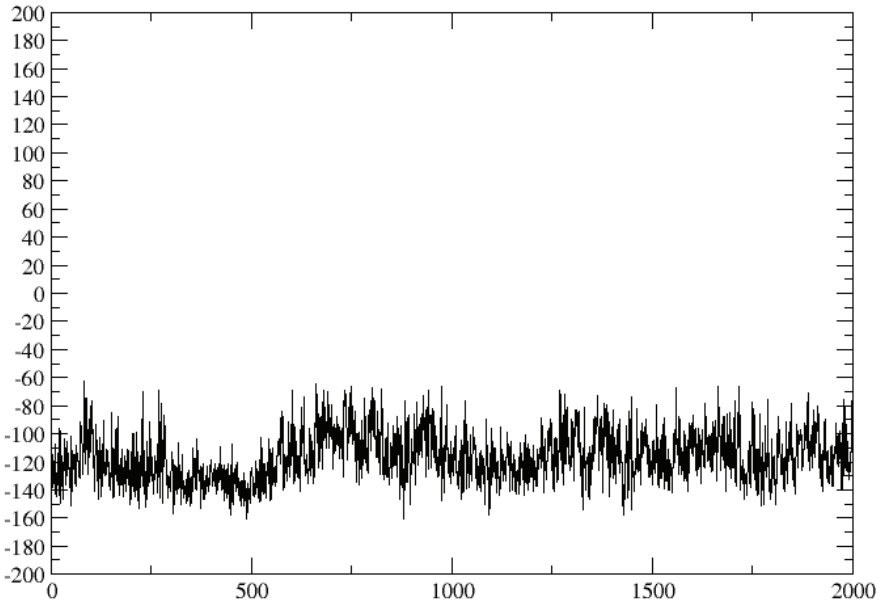



**Figure B.8. Phe33 Phi Dihedral Map**

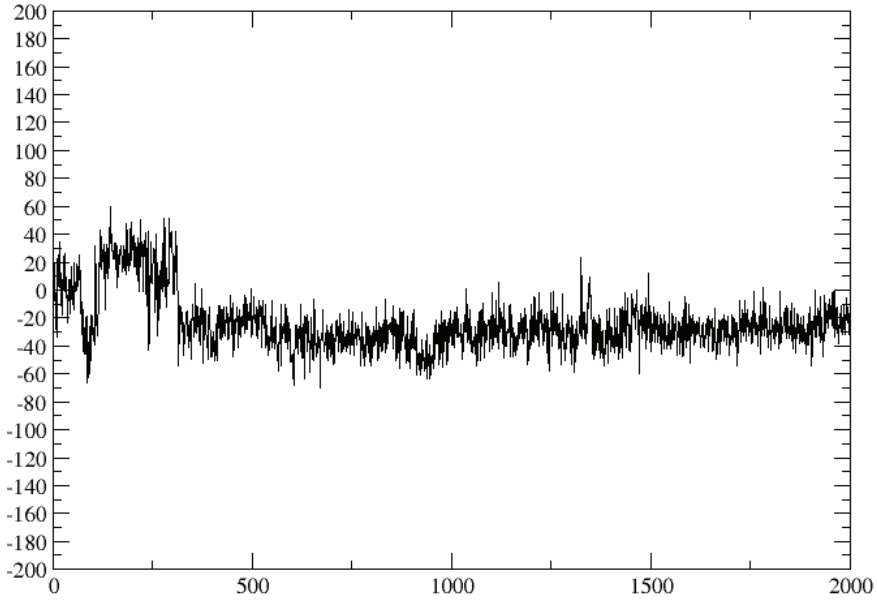



**Figure B.9. Phe33 Psi Dihedral Map**

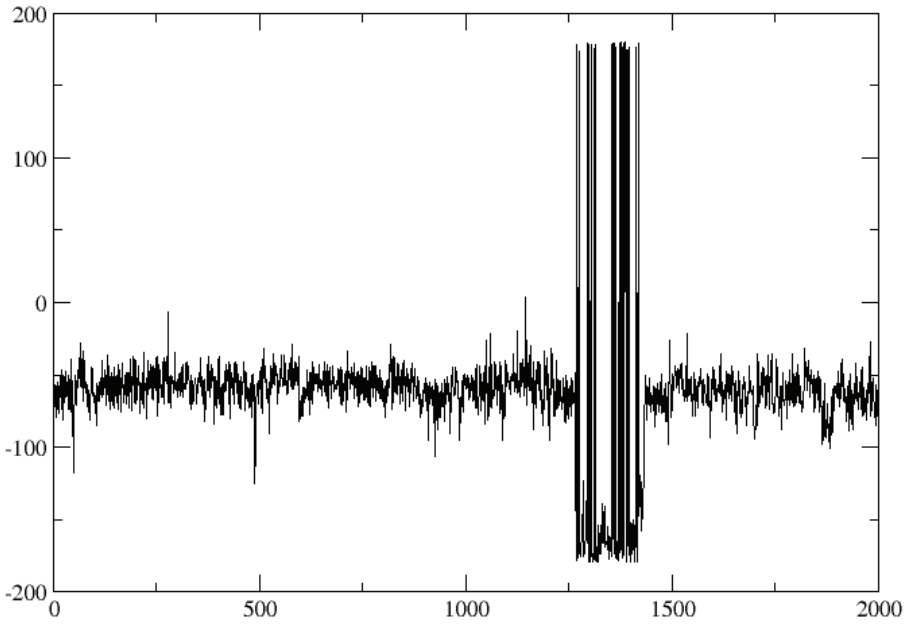



**Figure B.10. Phe33 Chi1 Dihedral Map**

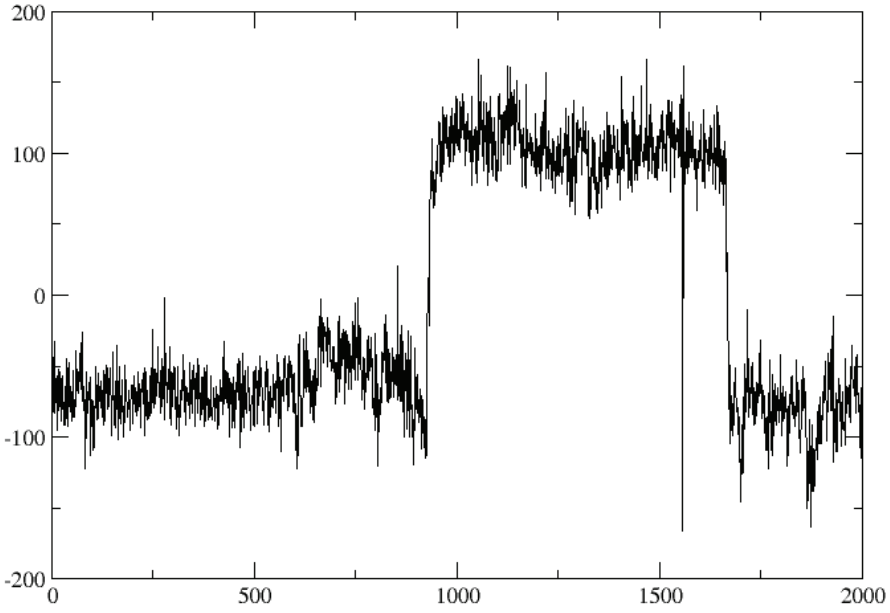



**Figure B.11. Phe33 Chi2 Dihedral Map**

### *B.3.2 Thr67 Dihedral Analysis*

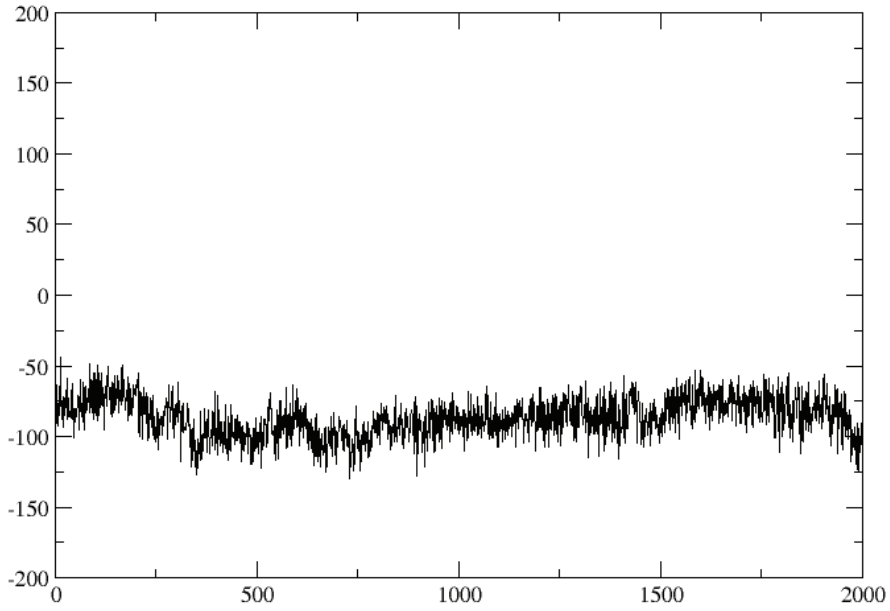



**Figure B.12. Thr67 Phi Dihedral Map**

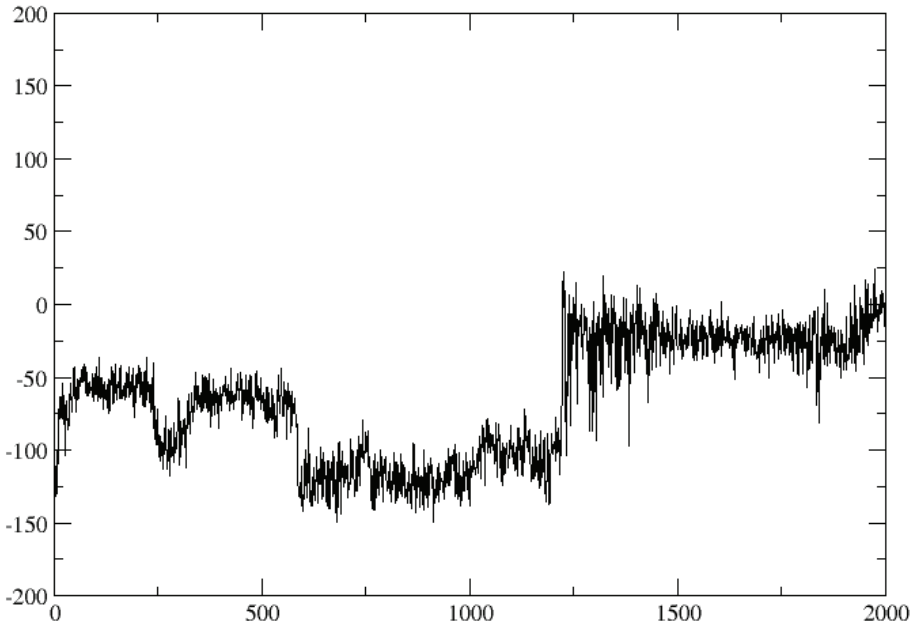



**Figure B.13. Thr67 Psi Dihedral Map**

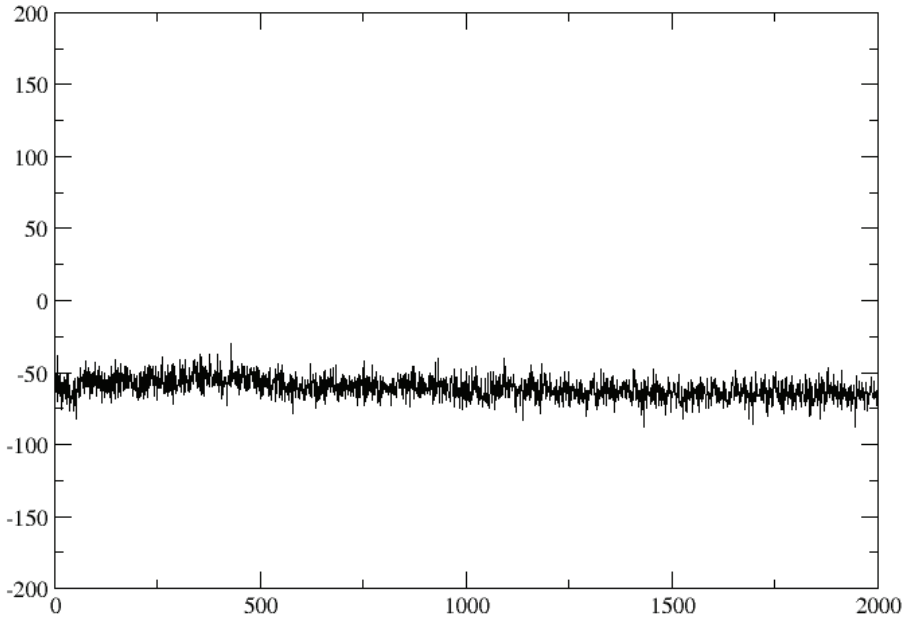



**Figure B.14. Thr67 Chi Dihedral Map**

### *B.3.3. Pro69 Dihedral Analysis*

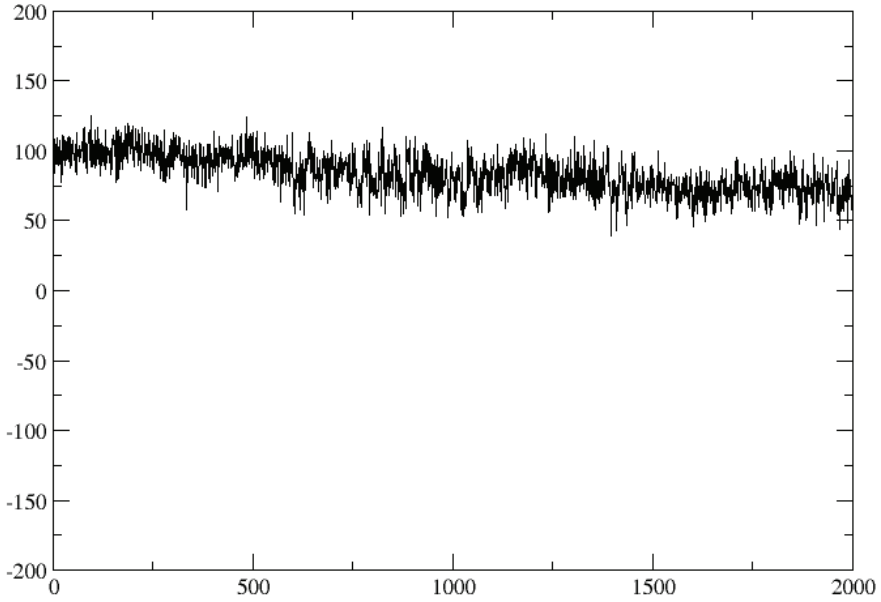



**Figure B.15. Pro69 Phi Dihedral Map**

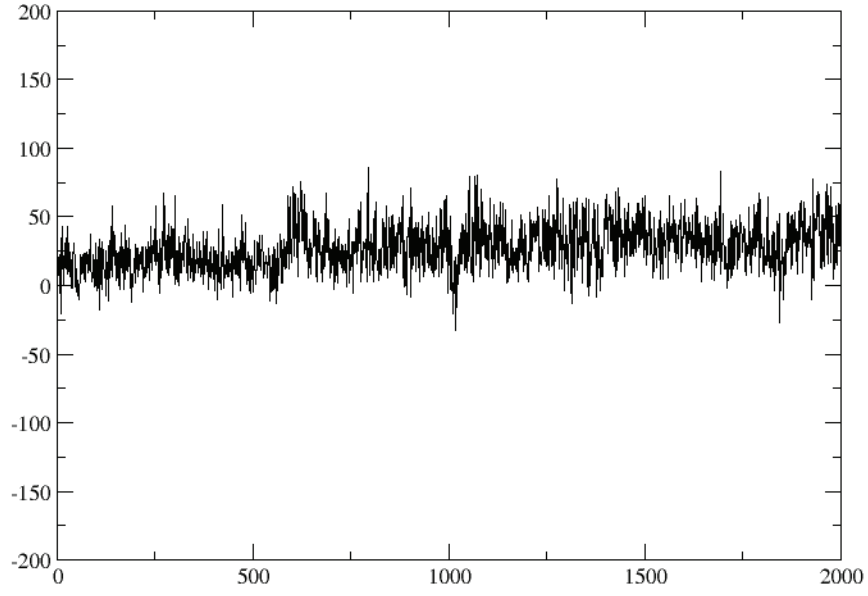



**Figure B.16. Pro69 Psi Dihedral Map**

### *B.3.4. Arg68 Dihedral Analysis*

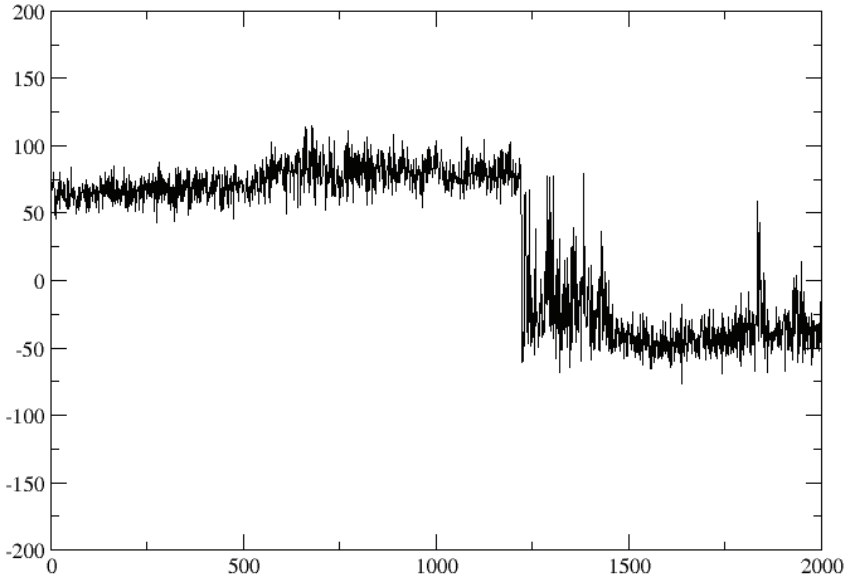



**Figure B.17. Arg68 Phi Dihedral Map**

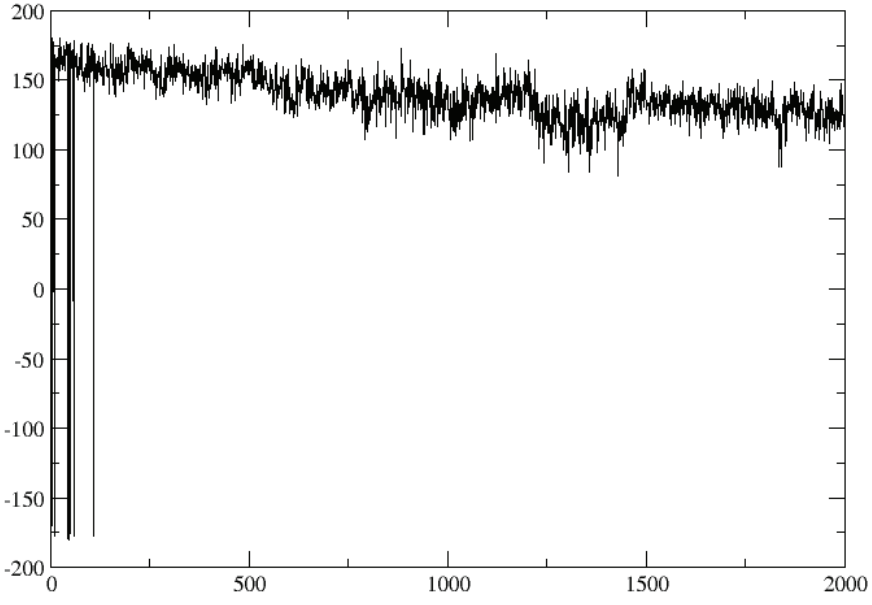



**Figure B.18. Arg68 Psi Dihedral Map**

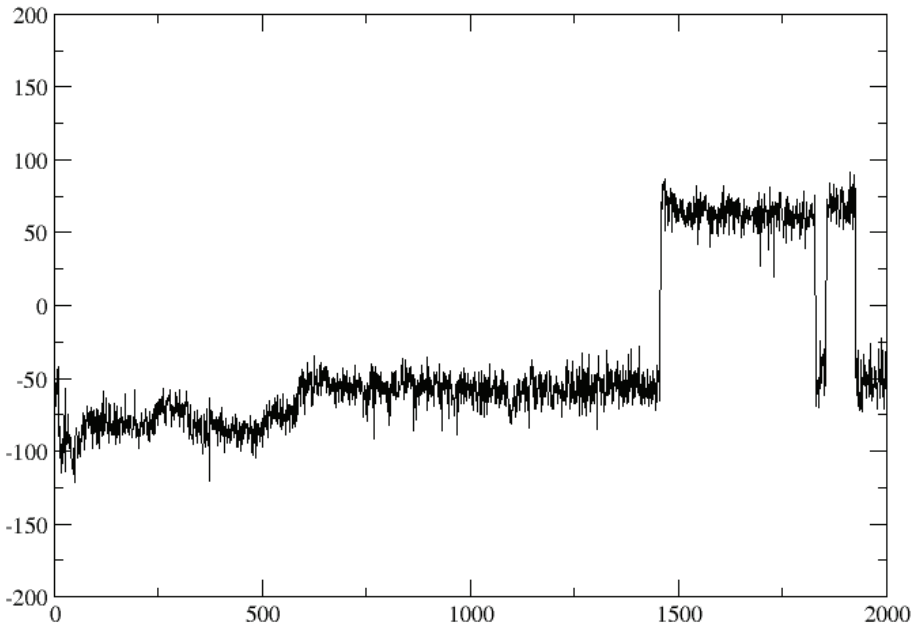



**Figure B.19. Arg68 Chi1 Dihedral Map**

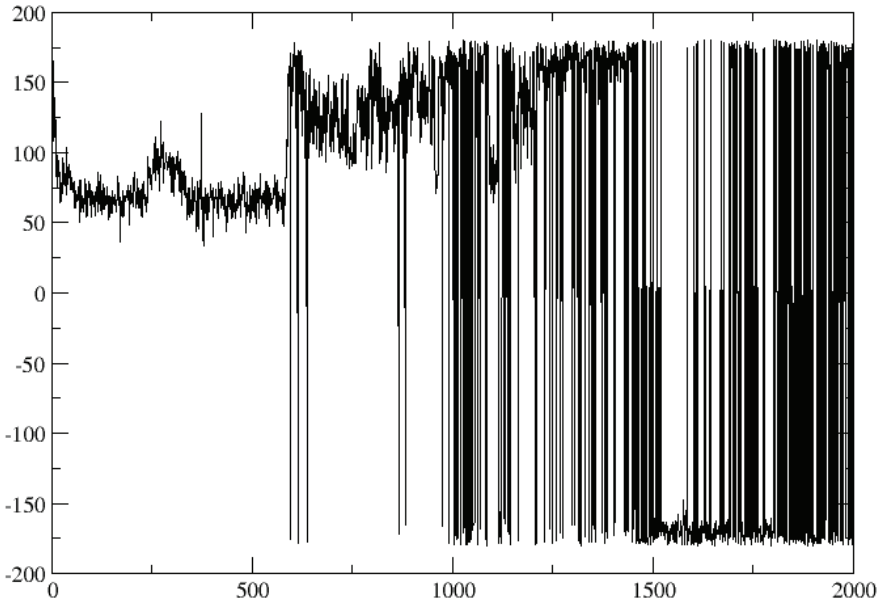



**Figure B.20. Arg68 Chi2 Dihedral Map**

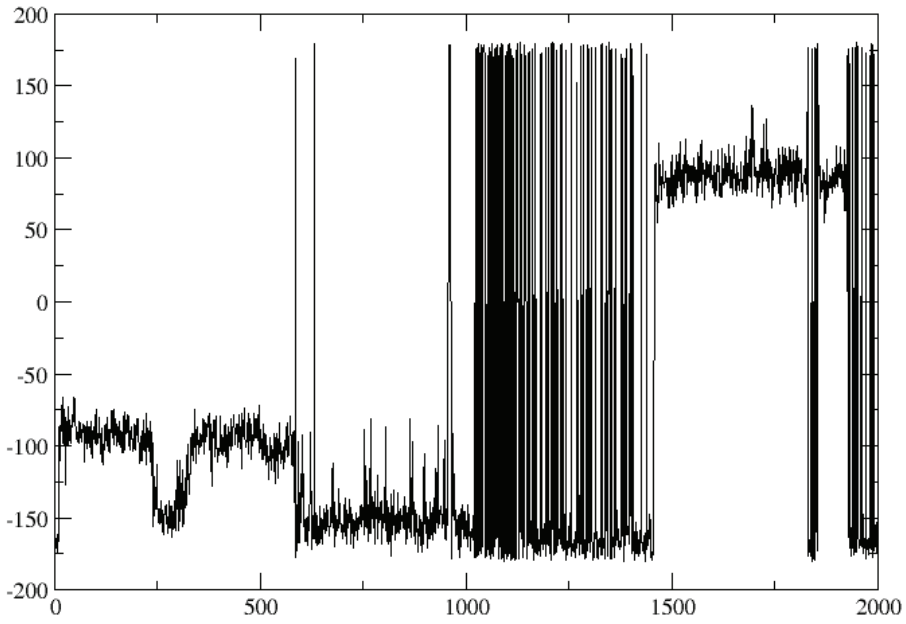



**Figure B.21. Arg68 Chi3 Dihedral Map**

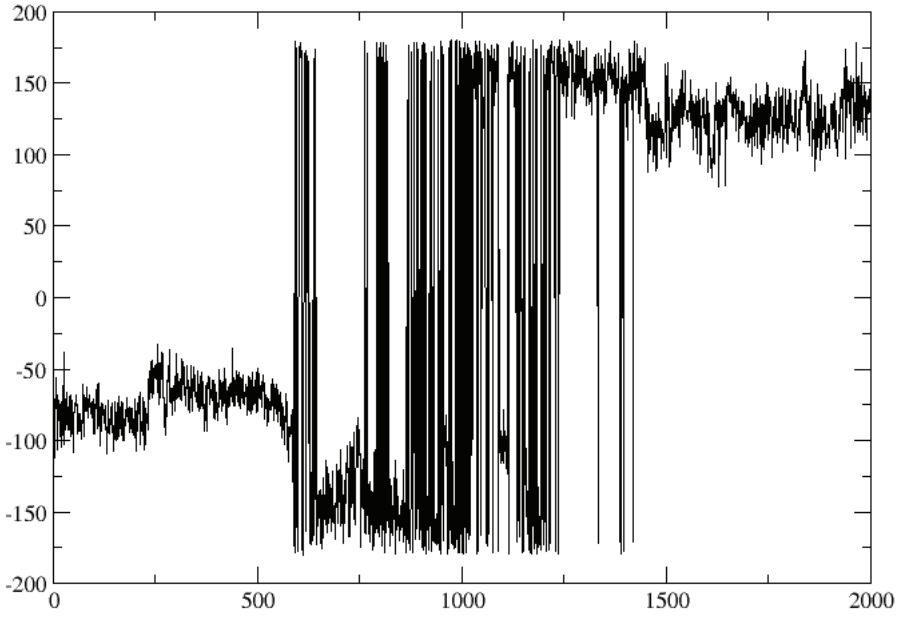



**Figure B.22. Arg68 Chi4 Dihedral Map**

## C. Chapter 3 Supplemental Tables and Figures

**Table C.1. DOCK Docking of the ACD Decoy Set**

| Docking Program / Validation Set | Grid Score | F-Score | PMF-Score | G-Score | D-Score | Chem-Score |
|---|---|---|---|---|---|---|
| **Grid Score** | --- | .748 | .155 | <.001 | <.001 | <.001 |
| **F-Score** | .237 | --- | .298 | .016 | .551 | .004 |
| **PMF-Score** | .975 | .154 | --- | <.001 | .006 | <.001 |
| **G-Score** | <.001 | .002 | <.001 | --- | .004 | .764 |
| **D-Score** | <.001 | .024 | <.001 | .008 | --- | .041 |
| **ChemScore** | .001 | .003 | <.001 | .265 | .043 | --- |

**Table C.2. DOCK Docking of the Schrödinger Decoy Set**

| Docking Program / Validation Set | Grid Score | F-Score | PMF-Score | G-Score | D-Score | Chem-Score |
|---|---|---|---|---|---|---|
| **Grid Score** | --- | .762 | .003 | .002 | .003 | .007 |
| **F-Score** | .161 | --- | .125 | .005 | .333 | <.001 |
| **PMF-Score** | <.001 | .774 | --- | <.001 | <.001 | <.001 |
| **G-Score** | .008 | .002 | <.001 | --- | .076 | .422 |
| **D-Score** | .056 | .053 | <.001 | .267 | --- | .085 |
| **ChemScore** | .047 | <.001 | <.001 | .467 | .165 | --- |

**Table C.3. DOCK Docking of the ZINC Decoy Set**

| Docking Program / Validation Set | Grid Score | F-Score | PMF-Score | G-Score | D-Score | Chem-Score |
|---|---|---|---|---|---|---|
| **Grid Score** | --- | .862 | .030 | <.001 | .002 | .001 |
| **F-Score** | .310 | --- | .119 | .002 | .515 | .001 |
| **PMF-Score** | <.001 | .665 | --- | <.001 | .001 | <.001 |
| **G-Score** | <.001 | .002 | <.001 | --- | .005 | .924 |
| **D-Score** | <.001 | .077 | <.001 | .005 | --- | .074 |
| **ChemScore** | .008 | .001 | <.001 | .911 | .162 | --- |

**Table C.4. FlexX Docking of the ACD Decoy Set**

| Docking Program / Validation Set | F-Score | PMF-Score | G-Score | D-Score | ChemScore |
|---|---|---|---|---|---|
| **F-Score** | --- | .537 | <.001 | .019 | <.001 |
| **PMF-Score** | .485 | --- | <.001 | .003 | <.001 |
| **G-Score** | .002 | <.001 | --- | .005 | .918 |
| **D-Score** | .017 | <.001 | .001 | --- | .010 |
| **ChemScore** | <.001 | <.001 | .331 | .033 | --- |

**Table C.5. FlexX Docking of the Schrödinger Decoy Set**

| Docking Program / Validation Set | F-Score | PMF-Score | G-Score | D-Score | ChemScore |
|---|---|---|---|---|---|
| **F-Score** | --- | .032 | <.001 | .011 | <.001 |
| **PMF-Score** | .076 | --- | <.001 | <.001 | <.001 |
| **G-Score** | .021 | <.001 | --- | .023 | .324 |
| **D-Score** | .030 | <.001 | .049 | --- | .001 |
| **ChemScore** | <.001 | <.001 | .184 | .028 | --- |

**Table C.6. FlexX Docking of the ZINC Decoy Set**

| Docking Program / Validation Set | F-Score | PMF-Score | G-Score | D-Score | ChemScore |
|---|---|---|---|---|---|
| **F-Score** | --- | .052 | <.001 | .017 | <.001 |
| **PMF-Score** | .051 | --- | <.001 | <.001 | <.001 |
| **G-Score** | .010 | <.001 | --- | .006 | .471 |
| **D-Score** | .042 | <.001 | .013 | --- | .001 |
| **ChemScore** | <.001 | <.001 | .224 | .028 | --- |

**Table C.7. Glide Docking of the ACD Decoy Set**

| Docking Program / Validation Set | Glide Score | F-Score | PMF-Score | G-Score | D-Score | Chem-Score |
|---|---|---|---|---|---|---|
| **Glide Score** | --- | .909 | .054 | <.001 | <.001 | <.001 |
| **F-Score** | .985 | --- | .027 | <.001 | <.001 | <.001 |
| **PMF-Score** | .142 | .317 | --- | <.001 | <.001 | <.001 |
| **G-Score** | <.001 | <.001 | <.001 | --- | .022 | <.001 |
| **D-Score** | <.001 | <.001 | <.001 | .031 | --- | .870 |
| **ChemScore** | .003 | <.001 | .002 | <.001 | .981 | --- |

**Table C.8. Glide Docking of the Schrödinger Decoy Set**

| Docking Program / Validation Set | Glide Score | F-Score | PMF-Score | G-Score | D-Score | Chem-Score |
|---|---|---|---|---|---|---|
| **Glide Score** | --- | .124 | .021 | <.001 | <.001 | <.001 |
| **F-Score** | .371 | --- | .140 | <.001 | <.001 | <.001 |
| **PMF-Score** | .089 | .529 | --- | <.001 | <.001 | <.001 |
| **G-Score** | <.001 | <.001 | <.001 | --- | .338 | .236 |
| **D-Score** | <.001 | <.001 | <.001 | .342 | --- | .966 |
| **ChemScore** | <.001 | <.001 | .005 | .625 | .811 | --- |

**Table C.9. GOLD Docking of the ZINC Decoy Set**

| Docking Program / Validation Set | GOLD Score | F-Score | PMF-Score | G-Score | D-Score | Chem-Score |
|---|---|---|---|---|---|---|
| **GOLD Score** | --- | .077 | .029 | <.001 | <.001 | <.001 |
| **F-Score** | .005 | --- | .863 | <.001 | <.001 | <.001 |
| **PMF-Score** | .007 | .708 | --- | <.001 | <.001 | <.001 |
| **G-Score** | <.001 | .002 | <.001 | --- | .958 | .024 |
| **D-Score** | <.001 | .002 | <.001 | .639 | --- | .010 |
| **ChemScore** | <.001 | <.001 | <.001 | .221 | .012 | --- |

**Table C.10. Surflex Docking of the ACD Decoy Set**

| Docking Program / Validation Set | Surflex Score | F-Score | PMF-Score | G-Score | D-Score | Chem-Score |
|---|---|---|---|---|---|---|
| **Surflex Score** | --- | .963 | .002 | <.001 | <.001 | <.001 |
| **F-Score** | .460 | --- | .008 | <.001 | <.001 | <.001 |
| **PMF-Score** | .003 | <.001 | --- | <.001 | <.001 | <.001 |
| **G-Score** | <.001 | <.001 | <.001 | --- | .001 | <.001 |
| **D-Score** | <.001 | <.001 | <.001 | | --- | .460 |
| **ChemScore** | <.001 | <.001 | <.001 | <.001 | .447 | --- |

**Table C.11. Surflex Docking of the ZINC Decoy Set**

| Docking Program / Validation Set | Surflex Score | F-Score | PMF-Score | G-Score | D-Score | Chem-Score |
|---|---|---|---|---|---|---|
| **Surflex Score** | --- | .499 | .172 | <.001 | <.001 | <.001 |
| **F-Score** | .475 | --- | .327 | <.001 | <.001 | <.001 |
| **PMF-Score** | .194 | .472 | --- | <.001 | <.001 | <.001 |
| **G-Score** | <.001 | <.001 | <.001 | --- | .006 | <.001 |
| **D-Score** | <.001 | <.001 | <.001 | .021 | --- | .249 |
| **ChemScore** | <.001 | <.001 | <.001 | <.001 | .258 | --- |

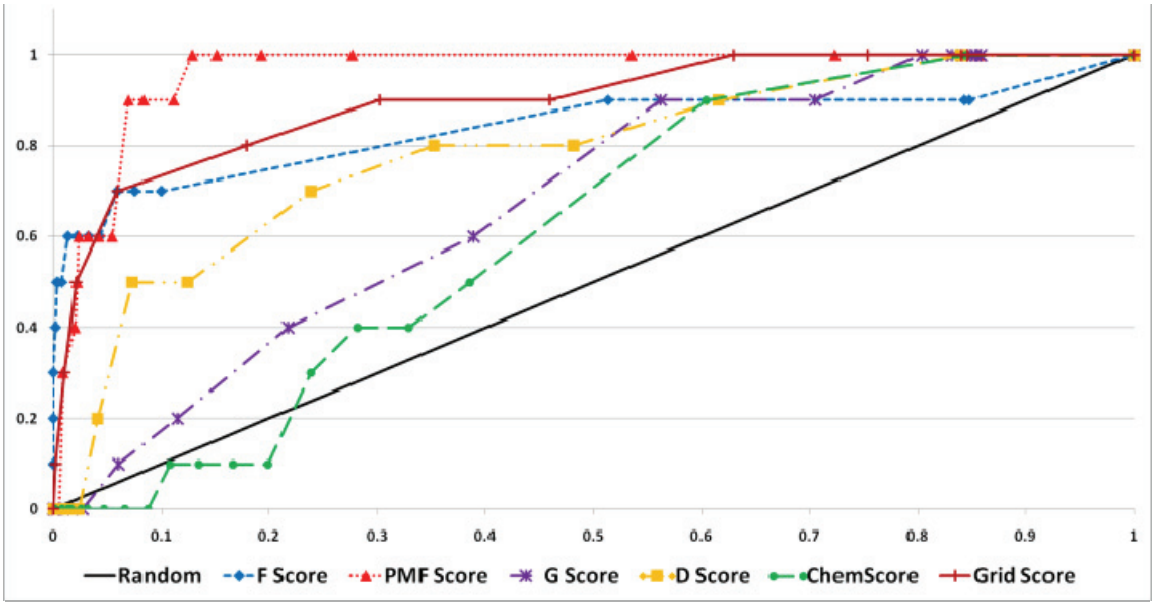



**Figure C.1. DOCK - ACD Decoy Set, ROC Curves**

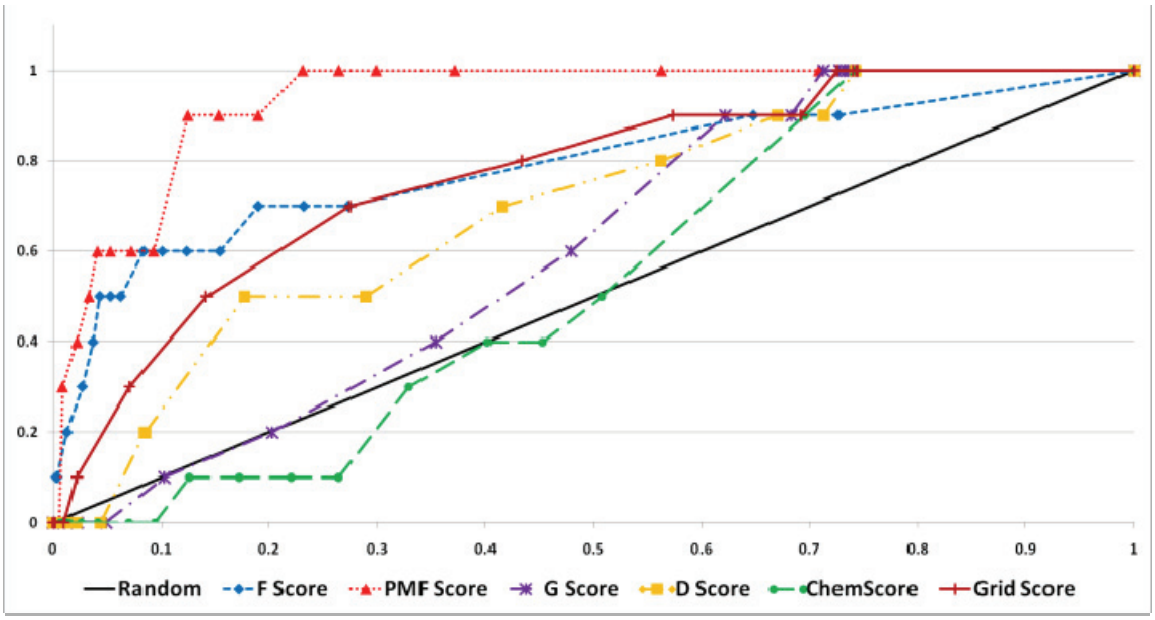



**Figure C.2. DOCK - Schrodinger Decoy Set, ROC Curves**

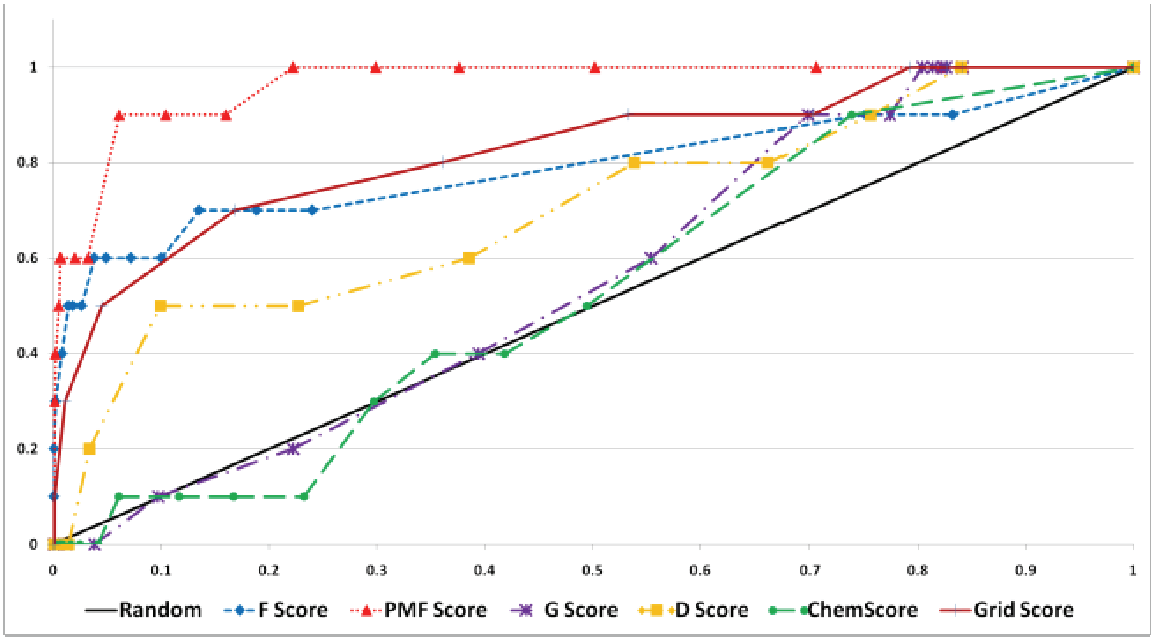



**Figure C.3. DOCK - ZINC Decoy Set, ROC Curves**

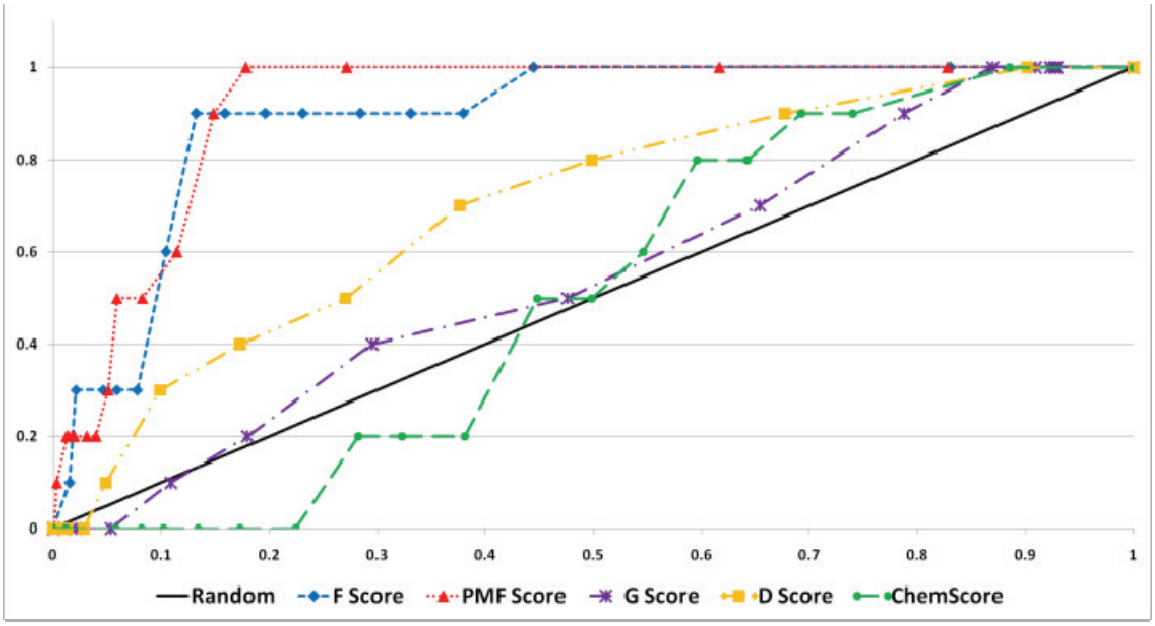



**Figure C.4. FlexX - ACD Decoy Set, ROC Curves**

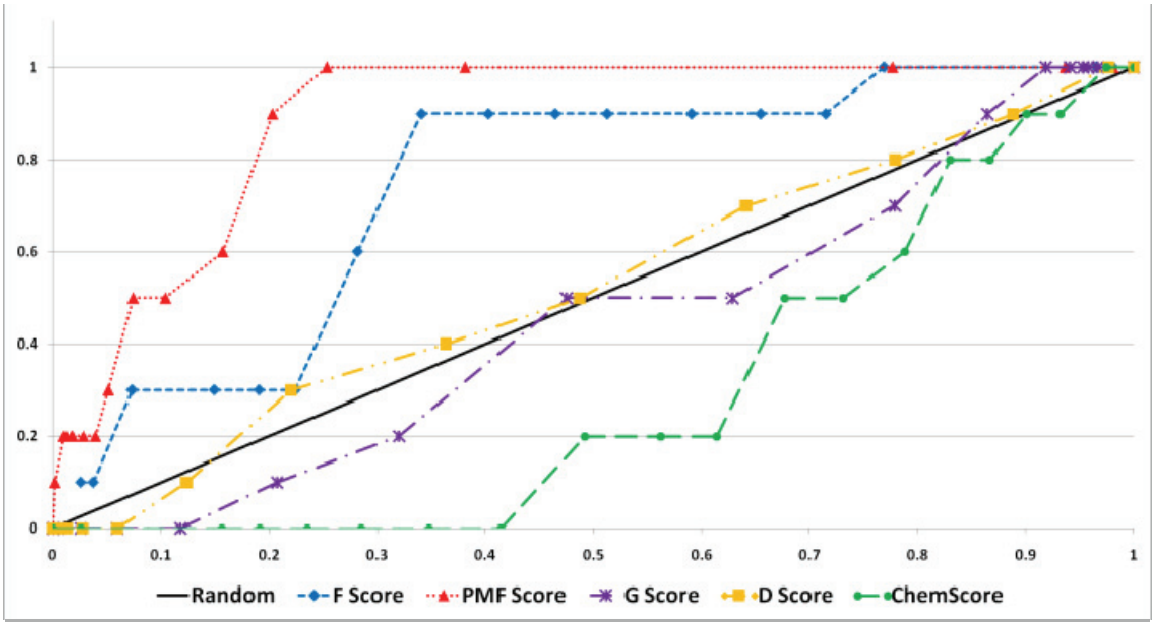



**Figure C.5. FlexX - Schrodinger Decoy Set, ROC Curves**

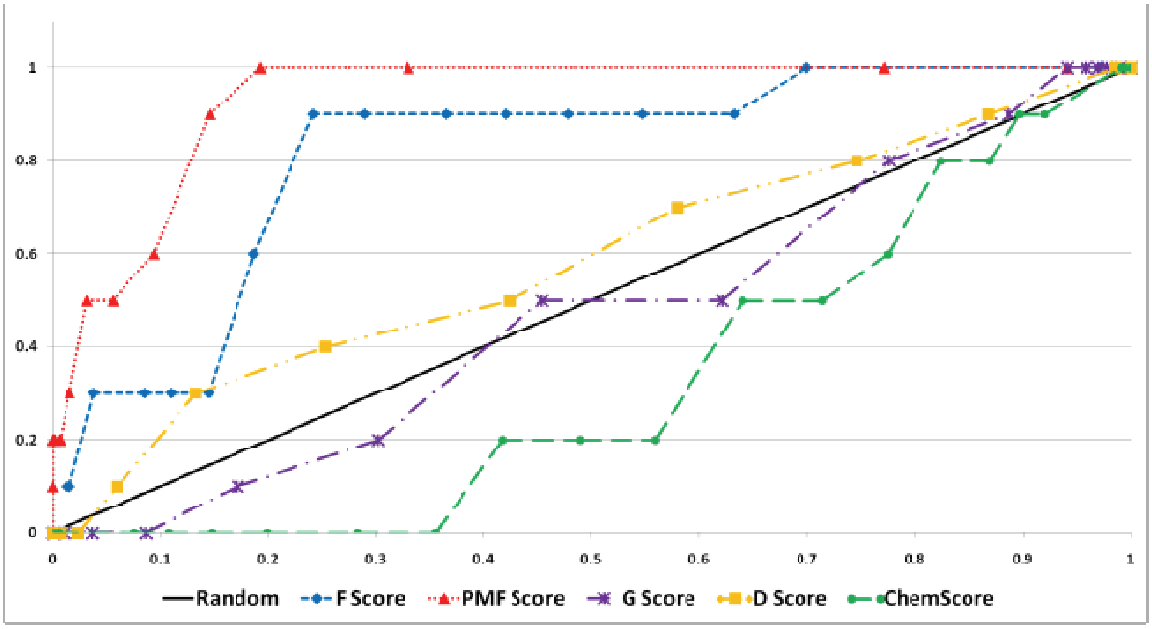



**Figure C.6. FlexX - ZINC Decoy Set, ROC Curves**

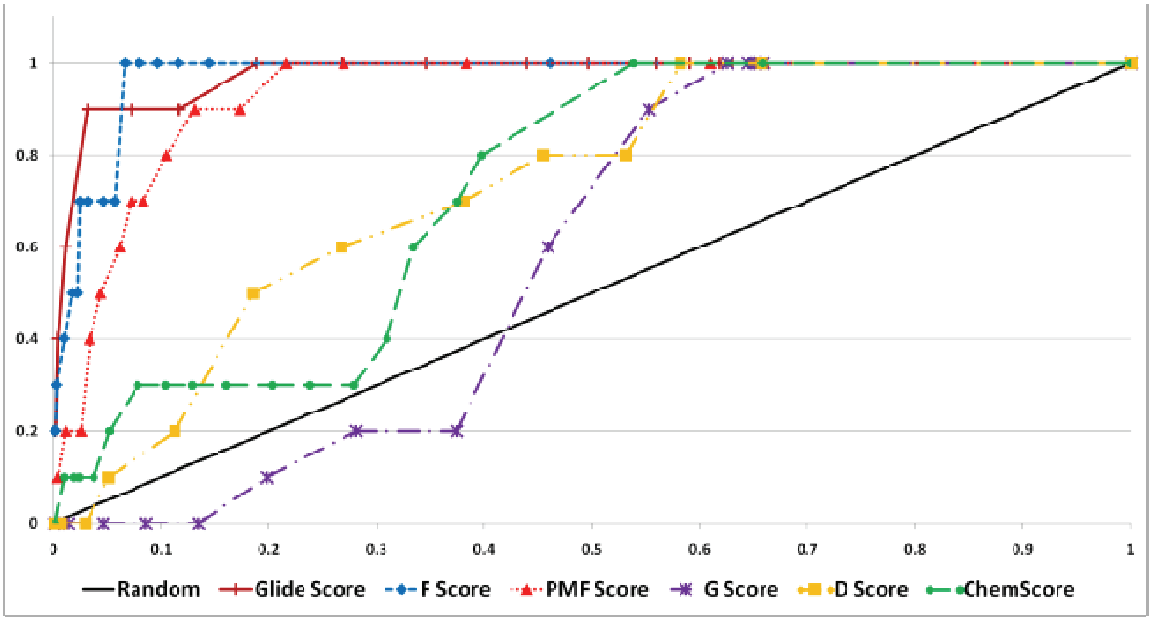



**Figure C.7. Glide - ACD Decoy Set, ROC Curves**

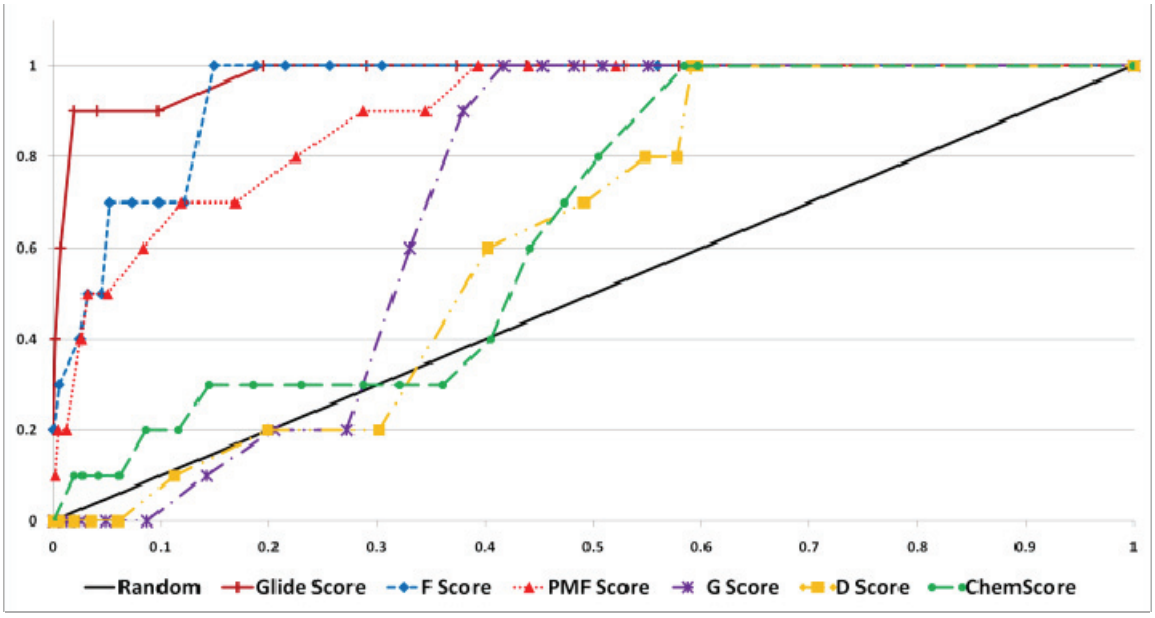



**Figure C.8. Glide - Schrodinger Decoy Set, ROC Curves**

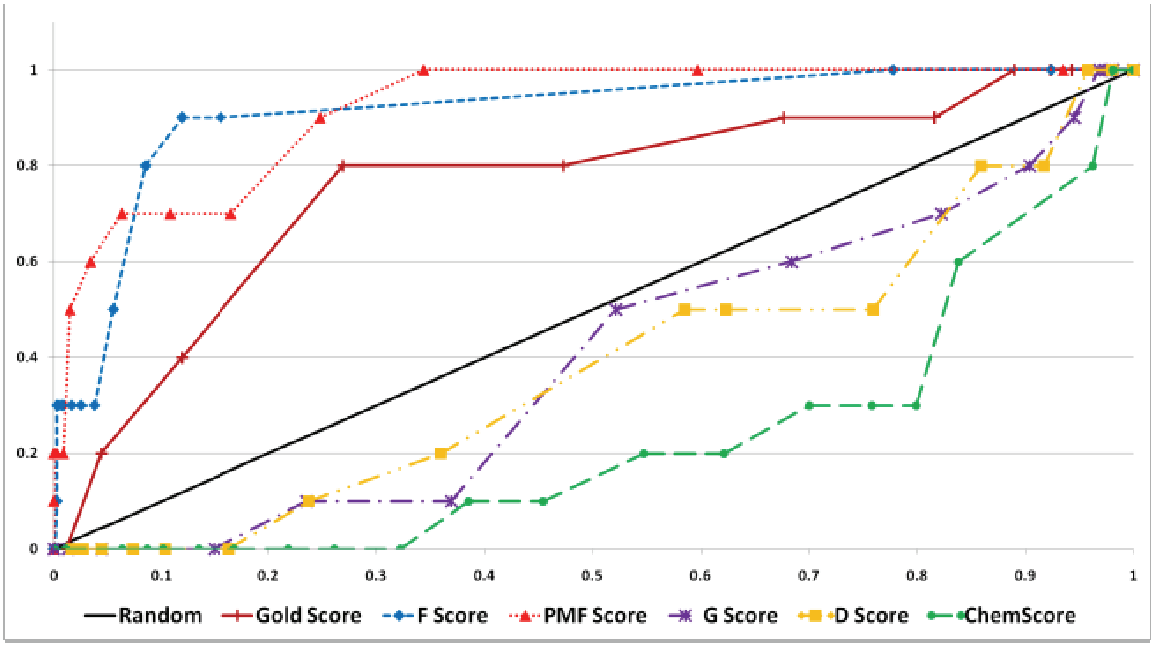



**Figure C.9. Gold - ZINC Decoy Set, ROC Curves**

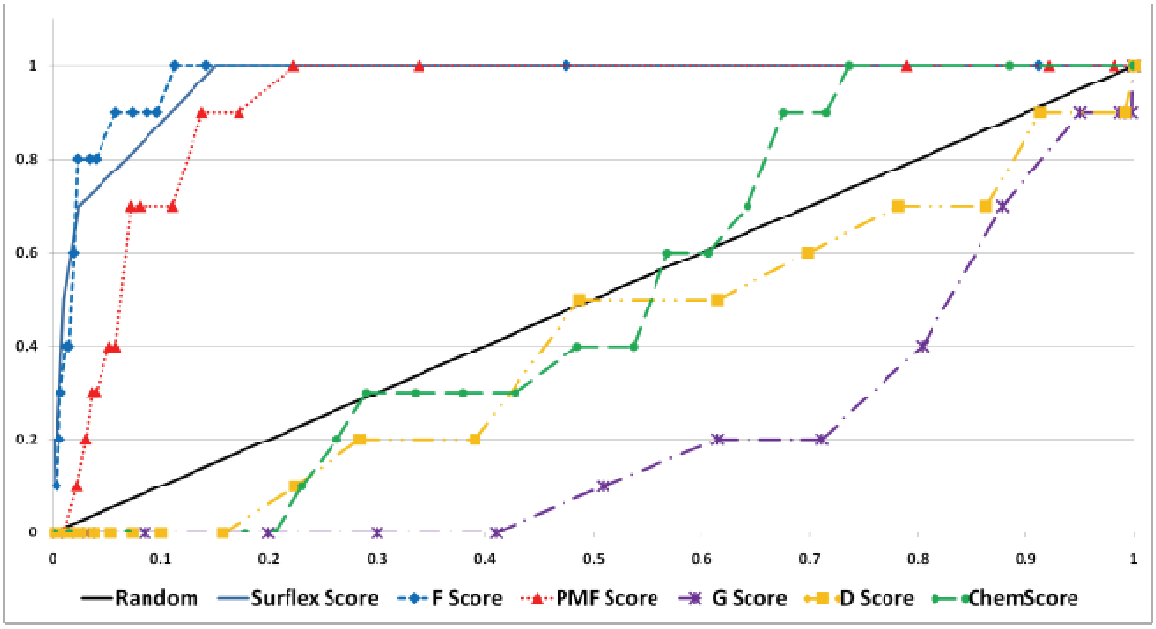



**Figure C.10. Surflex - ACD Decoy Set, ROC Curves**

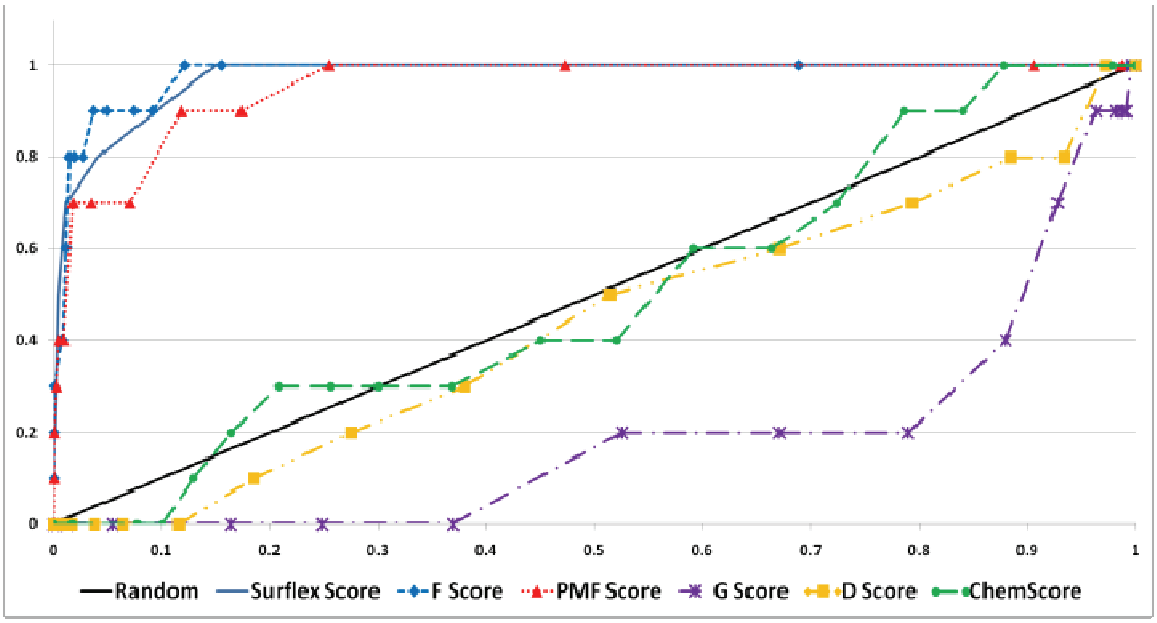



**Figure C.11. Surflex - ZINC Decoy Set, ROC Curves**

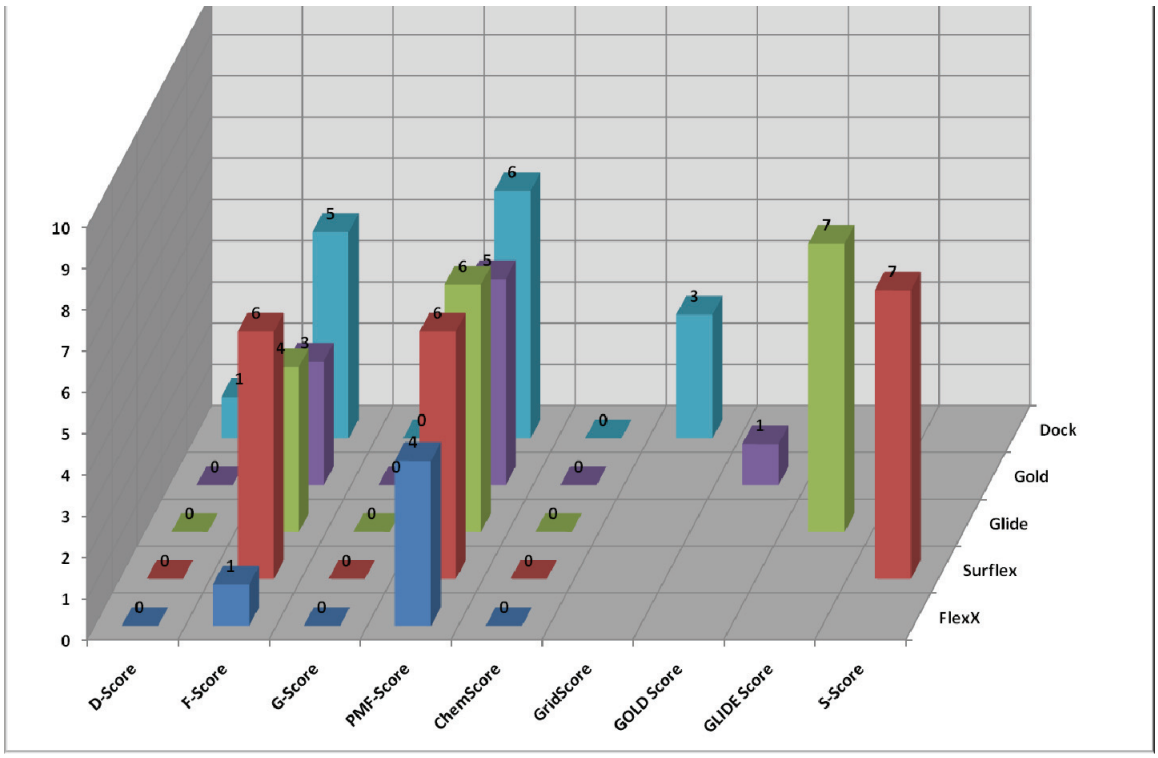



**Figure C.12. ZINC Decoy Set, Enrichment at 2%**

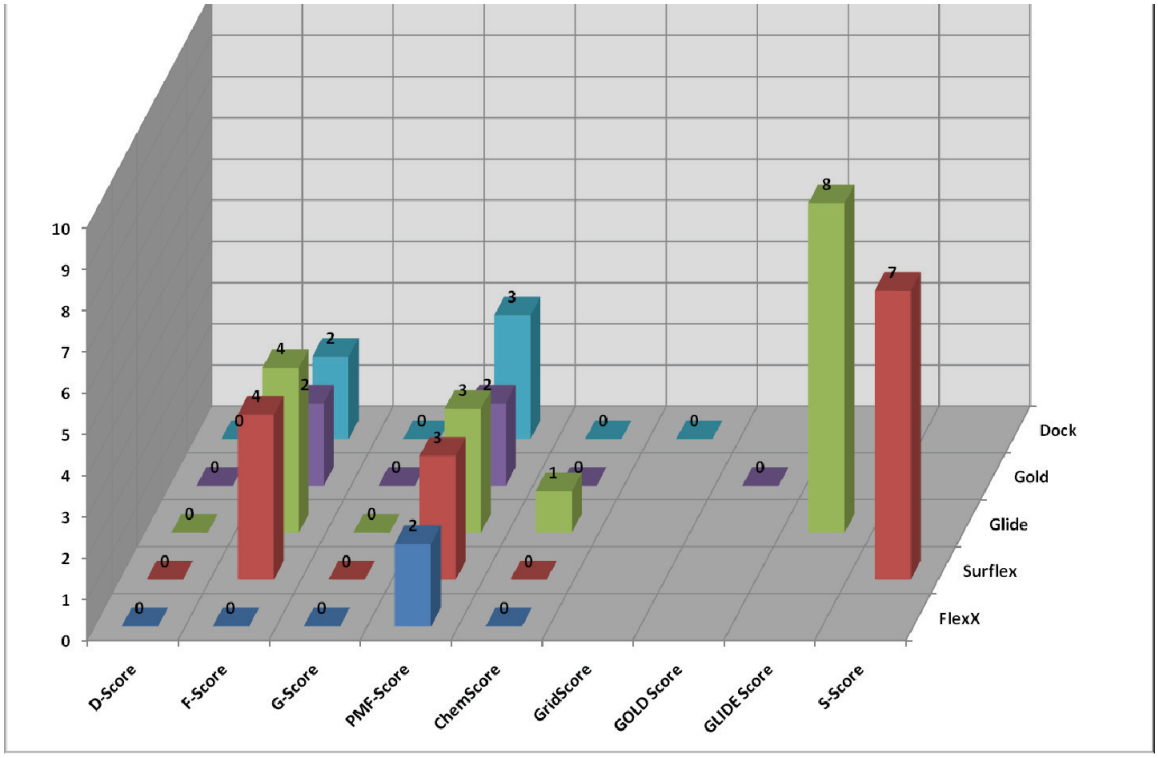



**Figure C.13. Schrodinger Decoy Set, Enrichment at 2%**

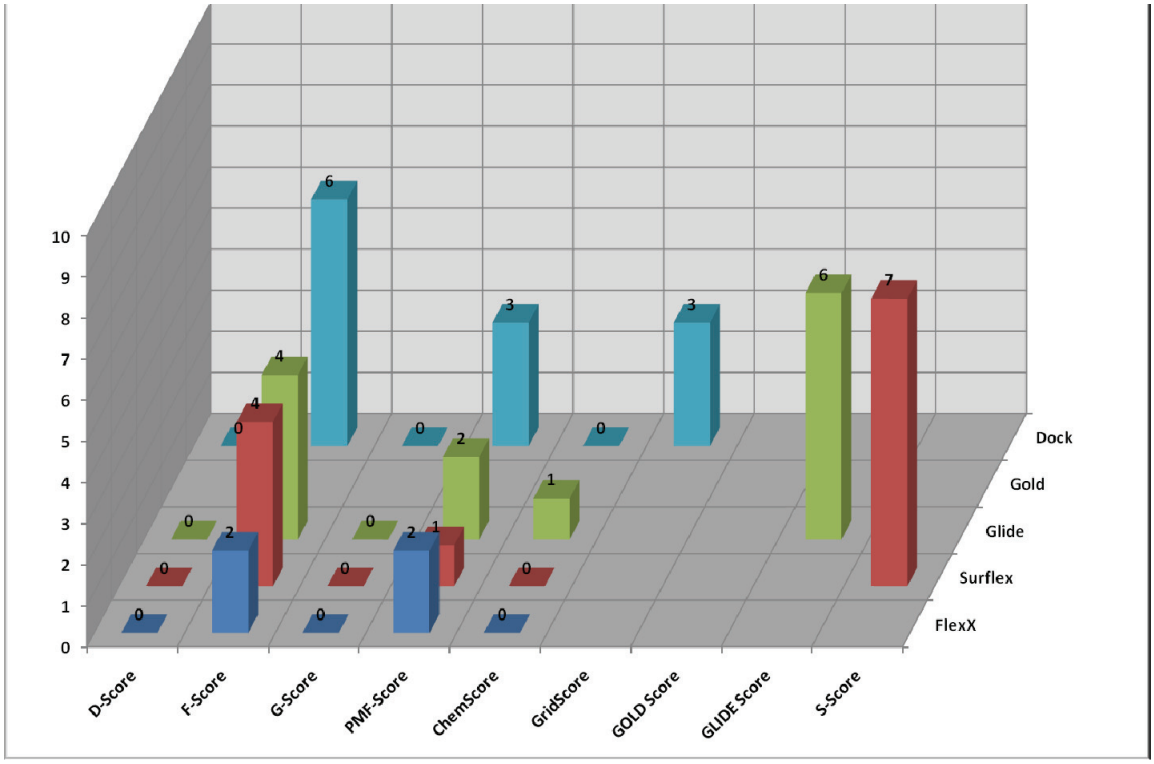



**Figure C.14. ACD Decoy Set, Enrichment at 2%**

# D. Chapter 4 Supplemental Material

## D.1 ZINC Databases Filtering Rules

```
#special flags
50.0 500.0 MOLWT
STRIPSALTS yes
0 10 CHIRALITY enumerate
ALLOWED_ATOMS C N O S P Cl F Br I H

# normal format is (min, max, name, SMARTS)

#rules
5 40 Non-Hydrogen_atoms [a,A]
2 40 carbons [#6]
1 20 N,O,S [#7,#8,#16]
0 1 Sulfonyl_halides S(=O)(=O)[Cl,Br]
0 1 Acid_halides [S,C](=[O,S])[F,Br,Cl,I]
0 1 Alkyl_halides [Br,Cl,I][CX4;CH,CH2]
0 0 Phosphenes cPc
0 0 Heptanes [CD1][CD2][CD2][CD2][CD2][CD2][CD2]
0 0 Perchlorates OCl(O)(O)(O)
0 7 Fluorines F
0 6 Cl,Br,I [Cl,Br,I]
0 0 Carbazides O=CN=[N+]=[N-]
0 0 Acid_anhydrides C(=O)OC(=O)
0 0 Peroxides OO
0 1 Iso(thio)cyanates N=C=[S,O]
0 1 Thiocyanates SC#N
0 0 Phosphoranes C=P
0 0 P/S_halides [P,S][Cl,Br,F,I]
#0 0 Carbodiimides N=C=N
0 0 Cyanohydrines N#CC[OH]
0 0 Carbazides O=CN=[N+]=[N-]
0 1 Sulfate_esters COS(=O)O[C,c]
0 1 Sulfonates COS(=O)(=O)[C,c]
0 0 Pentafluorophenyl_esters C(=O)Oc1c(F)c(F)c(F)c(F)c1(F)
0 0 Paranitrophenyl_esters C(=O)Oc1ccc(N(=O)=O)cc1
0 0 HOBt_esters C(=O)Onnn
0 0 Triflates OS(=O)(=O)C(F)(F)F
0 0 Lawesson's_reagents P(=S)(S)S
0 0 Phosphoramides NP(=O)(N)N
0 0 Aromatic_azides cN=[N+]=[N-]
0 2 Quaternary_C,Cl,I,P,S [C+,Cl+,I+,P+,S+]
```

0 2 Beta_carbonyl_quaternary_N C(=O)C[N+,n+]
0 2 Acylhydrazides [N;R0][N;R0]C(=O)
0 0 Chloramidines [Cl]C([C&R0])=N
0 0 Isonitriles [N+]#[C-]
0 0 Triacyloximes C(=O)N(C(=O))OC(=O)
0 0 Acyl_cyanides N#CC(=O)
0 0 Sulfonyl_cyanides S(=O)(=O)C#N
0 0 Cyanophosphonates P(OCC)(OCC)(=O)C#N
0 0 Azocyanamides [N;R0]=[N;R0]C#N
0 0 Azoalkanals [N;R0]=[N;R0]CC=O
0 2 (Thio)epoxides,aziridines C1[O,S,N]C1
0 2 Benzylic_quaternary_N cC[N+]
0 2 Thioesters C[O,S;R0][C;R0](=S)
0 3 Diand_Triphosphates P(=O)([OH])OP(=O)[OH]
0 2 Aminooxy(oxo) [#7]O[#6,#16]=O
0 2 nitros N(~[OD1])~[OD1]
0 2 Imines C=[N;R0]*
0 2 Acrylonitriles N#CC=C
0 2 Propenals C=CC(=O)[!#7;!#8]
0 4 Quaternary_N [n+,N+]

## D.2 Virtual Screen Round 1, All Compounds Selected for Screening

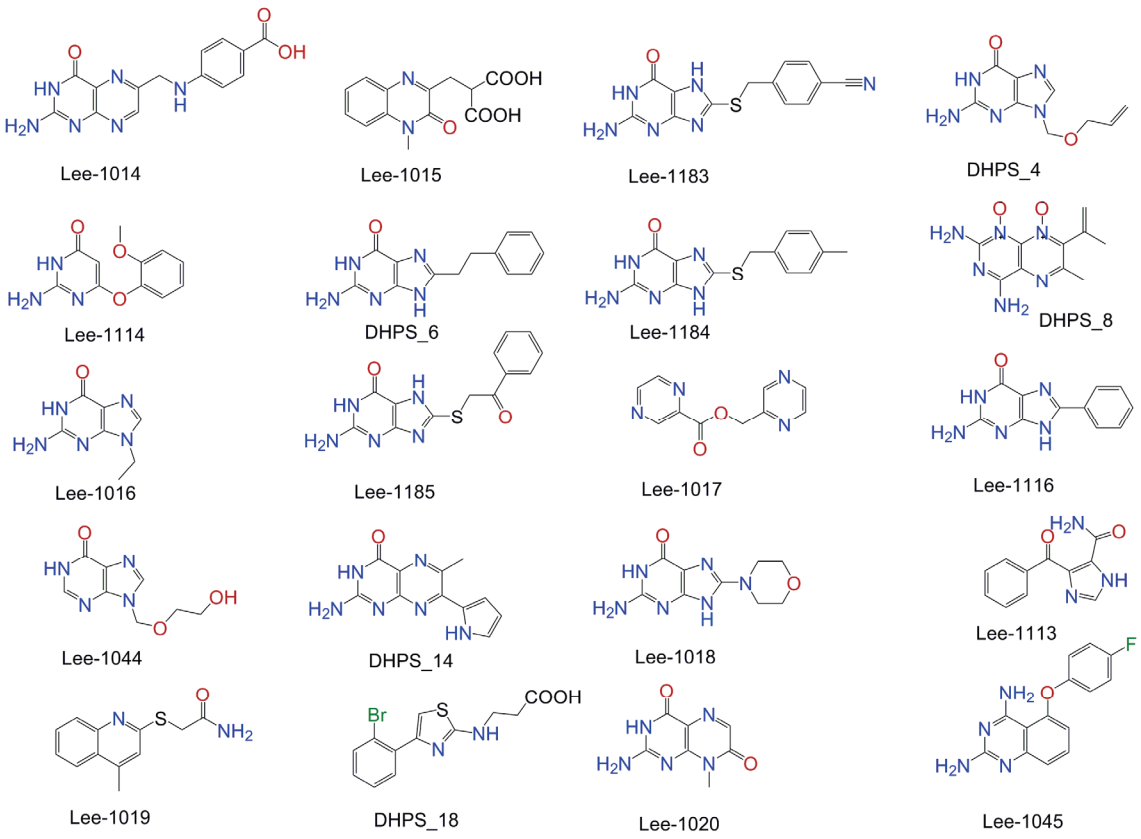



**Figure D.1. Virtual Screening, Round 1 Hits, Part 1**

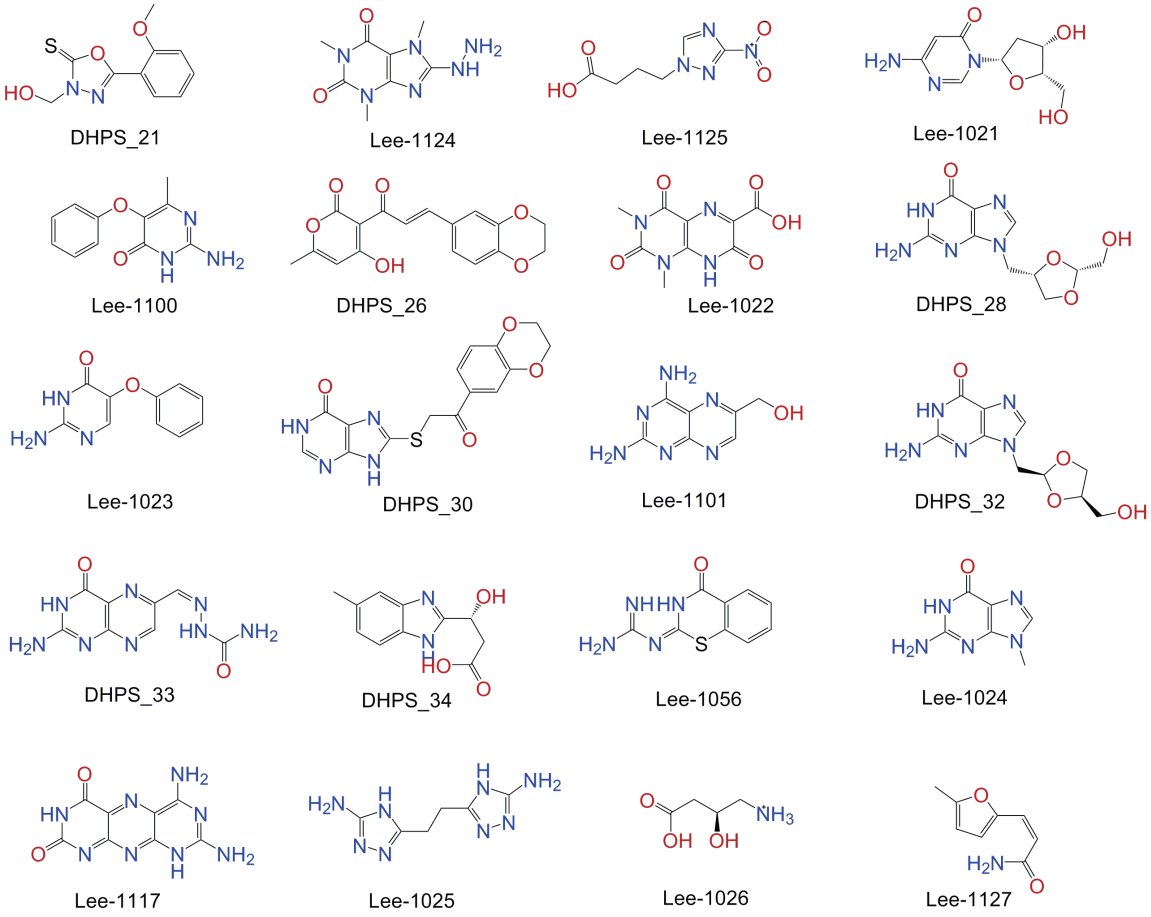



**Figure D.2. Virtual Screening, Round 1 Hits, Part 2**

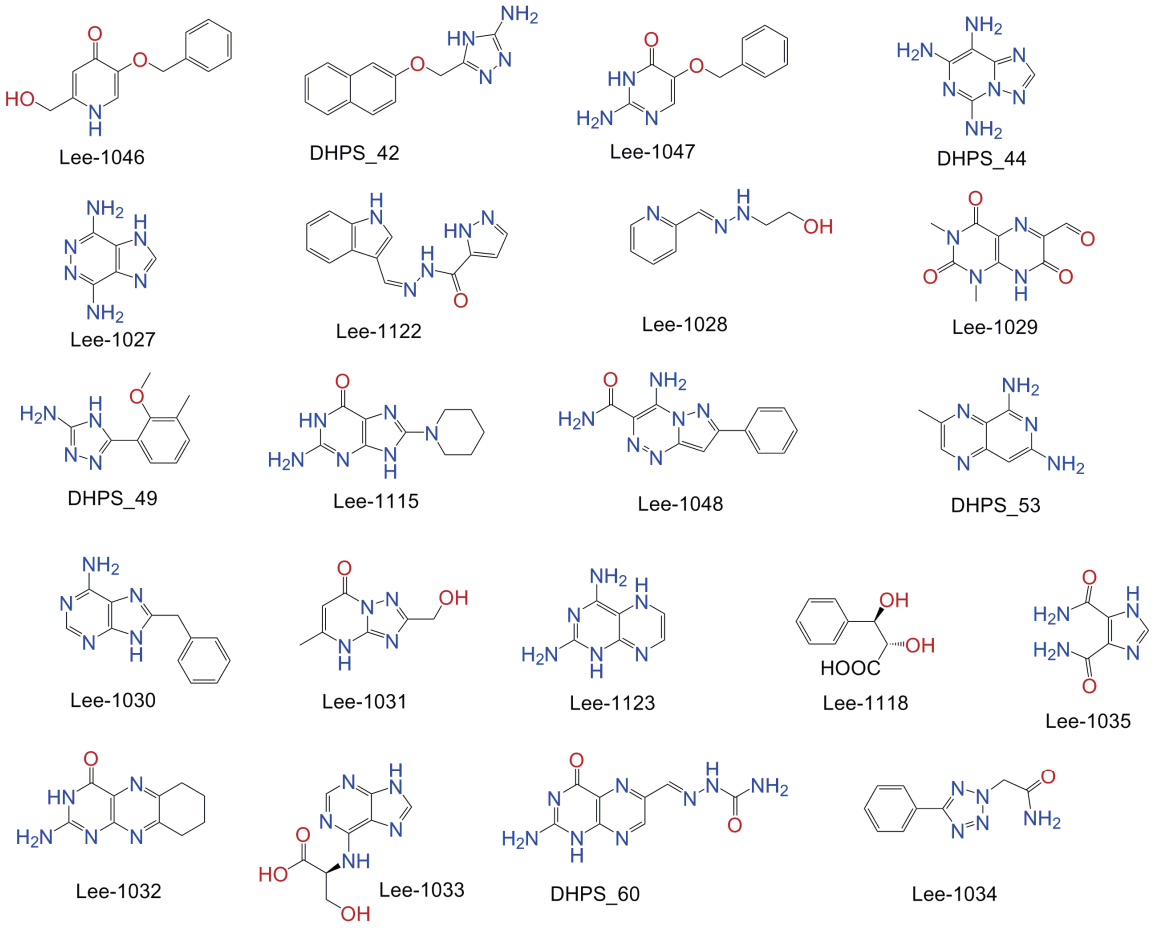



**Figure D.3. Virtual Screening, Round 1 Hits, Part 3**

## D.3 Virtual Screen Round 2, All Compounds Selected for Screening

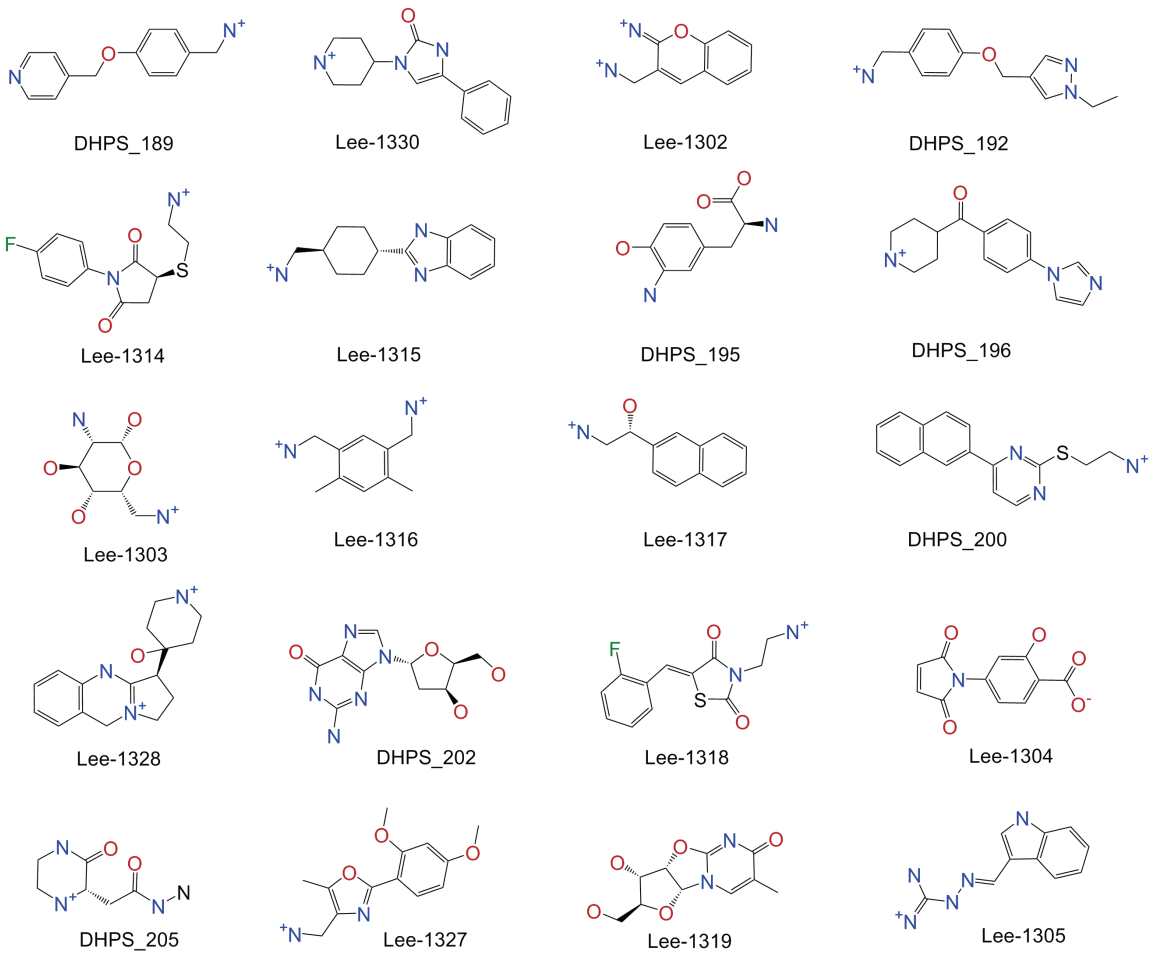



**Figure D.4. Virtual Screening, Round 2 Hits, Part 1**

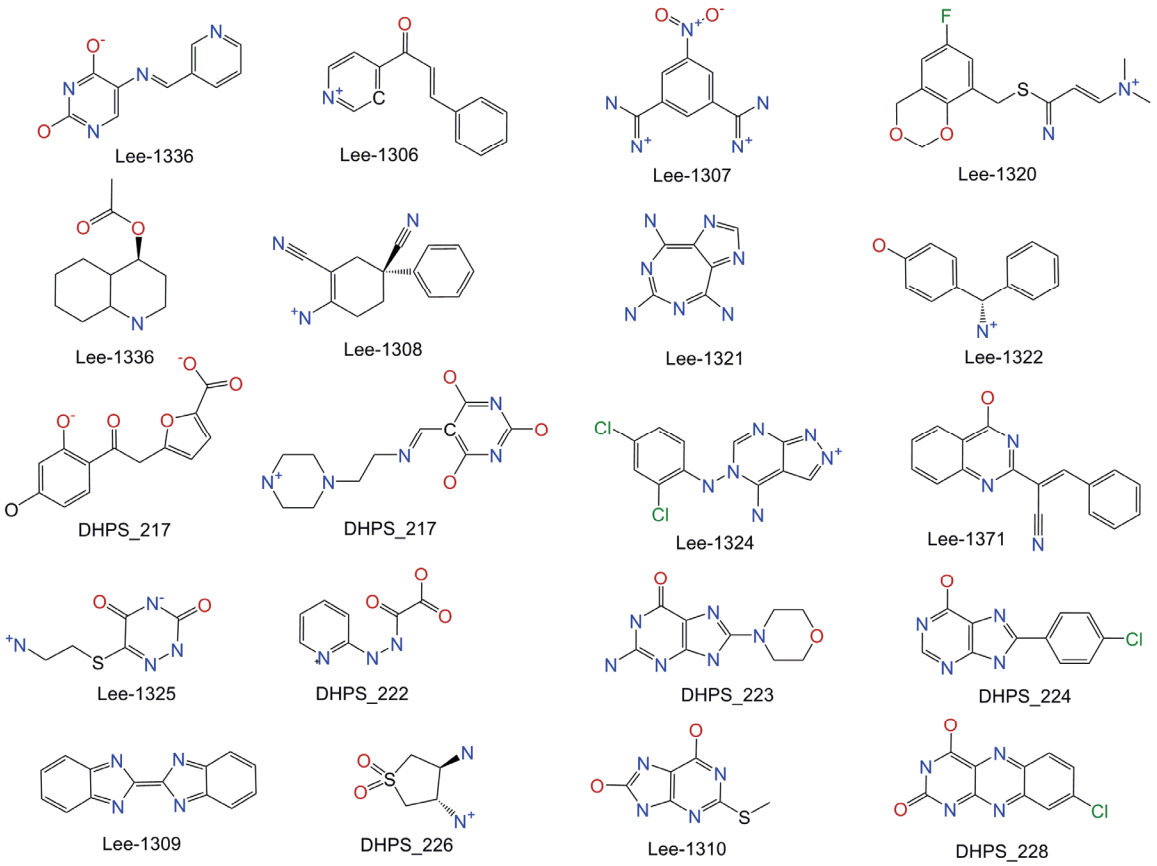



**Figure D.5. Virtual Screening, Round 2 Hits, Part 2**

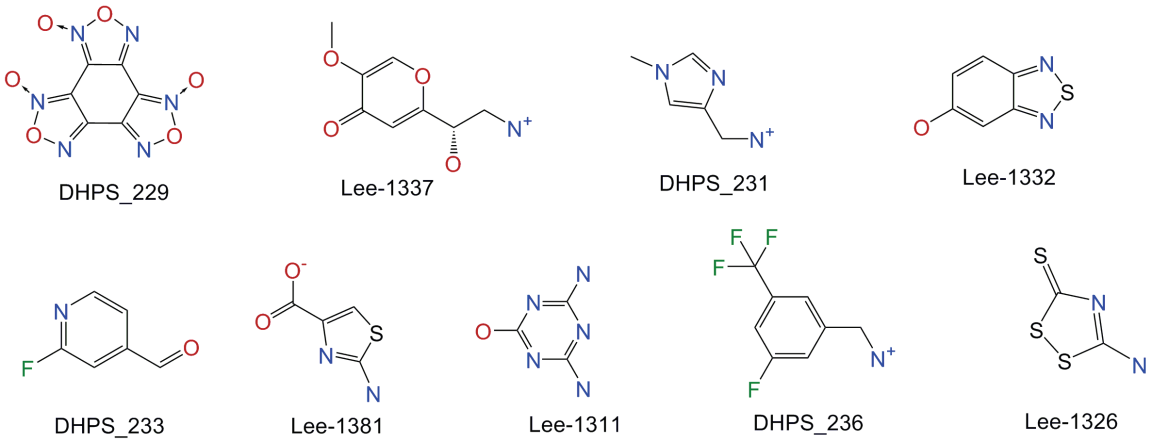



**Figure D.6. Virtual Screening, Round 2 Hits, Part 3**

## VITA

Kirk Edward Hevener was born in Akron, Ohio to Eugene and Ellen Hevener on March 7th, 1973. After graduating from Stow-Munroe Falls High School in 1991, he joined the U.S. Navy where he served as a Hospital Corpsman and Pharmacy Technician for 6 years prior to his honorable discharge. In 1997, upon completion of his military service, he relocated from southern California to Nashville, TN where he enrolled at Tennessee State University. He received a Bachelor of Science in Chemistry from TSU in December, 2005. He entered pharmacy school at the University of Tennessee in 2001, enrolling in the dual PharmD/PhD program, working toward both a pharmacy doctorate and a Medicinal Chemistry doctorate. He completed his pharmacy doctorate in May 2005 and his Medicinal Chemistry doctorate in December 2008.