Improvements in Molecular Mechanics Sampling and Energy Models

Joseph Bylund

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2013

ABSTRACT

Improvements in Molecular Mechanics Sampling and Energy Models

Joseph Bylund

The process of bringing drugs to market continues to be a slow and expensive affair. And despite recent advances in technology, the cost both in monetary terms and in terms of time between target identification and arrival of a new drug on the market continues to increase.

High throughput screening is a first step in the direction of being able to test a large number of possible bioactive compounds very quickly. However the space of possible small molecules is limitless, and high throughput screening is limited both by the size of available libraries and the cost of running such a large number of experiments.

Computational drug design, or computer assisted drug design offers a possible way of addressing some of the shortfalls of conventional high throughput screening. Using computational methods it is possible to estimate parameters such as binding affinity, of any small molecule, even those not currently present in any small molecule library, without having to first invest in the possibly slow and expensive process of finding a synthesis. Computational methods can be used to screen similar molecules, or mutations in small molecule space, seeking to increase binding affinity to the protein target, and thereby efficacy, while simultaneously minimizing binding affinity to other proteins, decreasing cross reactivity, and reducing toxicity and harmful side effects.

Computational biology methods of drug research can be broadly classified in a number of different ways. However; one of the most common classifications is along the lines of the methods used to identify possible drug compounds and later optimize those leads. The first broad category is the informatics or artificial intelligence based approaches. In these approaches artificial intelligence methods such as neural networks, support vector machines and qualitative structure-activity relationships (QSAR) are used to identify chemical or

structural properties that contribute heavily to binding affinity. Ligand based approaches are very useful when a large number of known binders are known for a specific family of proteins. In this case the ligands cluster together in some sort of chemical space and new compounds which occupy a similar chemical space are likely to also bind strongly with the protein of interest. The class explored in this thesis is the diverse class known as structural methods. These methods in the most general sense make use of a sampling method to sample a number of protein, or protein-small-molecule interaction conformations and an energy model or scoring function to measure dimensions which would be very difficult and or expensive to measure experimentally.

In this thesis a number of different sampling methods which are applicable to different questions in computational biology are presented. Additionally an improved algorithm for evaluating implicit solvent effects is presented, and a number of improvements in performance, reliability and utility of the molecular mechanics program used are discussed.

Table of Contents

1 Introduction **1**

1.1 Drug Development 1

1.1.1 Costs of Drug Development 1

1.1.2 Computer Assisted Drug Design 4

1.1.2.1 Hit Identification 4

1.1.2.2 Hit-to-Lead Optimization 9

1.1.2.3 Lead Optimization 10

1.2 Sampling Algorithms 11

1.2.1 Minimization 11

1.2.2 Monte-Carlo Sampling 12

1.2.3 Analytic Loop Closure 12

1.2.4 Random Tweak 13

1.2.5 Cyclic Coordinate Descent 14

1.2.6 Rotamer Assembly 14

1.3 Molecular Modeling 15

1.4 Energy Functions 17

1.4.1 The General Form of the Energy Model 18

1.4.2 Molecular Surfaces 19

1.4.3 Solvent Models 20

2 Cell Based Implicit Solvent **23**

2.1 Introduction 23

2.2 Methods . . . 26
2.2.1 Energy Model . . . 26
2.2.2 Data Sets . . . 27
2.2.3 Structure Preparation . . . 27
2.2.4 Grid-Based Spatial Indexing . . . 28
2.2.5 Experiments . . . 29
2.3 Results . . . 30
2.4 Discussion . . . 33

3 Computational Mutation Scanning **37**
3.1 Introduction . . . 37
3.1.1 Entropy-Enthalpy Compensation . . . 39
3.2 Methods . . . 39
3.2.1 General Mutation Screening . . . 39
3.2.2 Alanine Scanning Experiments . . . 40
3.3 Results . . . 41
3.4 Discussion . . . 55
3.4.1 Side Chain Prediction Accuracy . . . 55
3.4.2 Energetic Correlation with Experimental Data . . . 55
3.4.3 Future Directions . . . 56

4 Prediction of P450 Sites of Metabolism **57**
4.1 Introduction . . . 57
4.2 Methods . . . 59
4.2.1 Docking . . . 61
4.2.2 Monte Carlo Minimization Refinement . . . 62
4.2.3 Evaluation . . . 71
4.3 Results . . . 75
4.4 Discussion . . . 82
4.4.1 Significance of Sampling Stages . . . 82
4.4.2 Induced Fit Effects . . . 83

4.4.3 Balancing Structural Contribution and Reactivity 83

5 PLOP Improvements **85**

5.1 Regression Testing 85

5.2 Small Molecule Library 87

5.3 Knowledge Based Backbone Dihedral Penalty 89

Bibliography **90**

List of Figures

1.1 The rate at which new structures is deposited into the PDB over the last two decades. Due to a variety of improvements in the field of crystallography this rate has been steadily increasing. 6

1.2 The HIV protease inhibitor, nelfinavir, marketed under the name Viracept was originally identified using a computational docking screen. It has a very high binding affinity for its target protein, 2 nM. Here it is shown crystallized with multidrug variant (ACT) (V82T/I84V) of HIV-1 protease, PDBID 3EL5. (b) generated with Visual Molecular Dynamics [Humphrey *et al.*, 1996] and [POVRAY 3.6, 2004]. 8

1.3 To an extent it is always possible to either increase accuracy or decrease running time, or the cost of an experiment. New scientific methods should allow one to increase accuracy while not spending additional time. 16

1.4 Here energy is represented as a function of the two principal components of the protein conformation, in both cases the approximate funnel shape of the energy surface about the native conformation is very apparent. (a) shows an energy surface without any solvent effects. (b) represents a surface with the effect of hydration, though this surface appears smooth relative the dry surface in reality all energy landscapes of larger proteins contain many local minima. Figure from [Chaplin, 2013] used with authors permission. 17

1.5 (a) The Van der Waals surface of Nelfinavir, defined by the surface of the volume excluded by the VDW radii of the atoms in the structure. (b) The molecular surface, defined as the surface of the volume excluded from a probe of 1.4 angstroms (the radius of a water molecule). Both surfaces are enclosed by an approximate solvent accesible surface, which is defined as the surface traced by rolling a spherical probe over the VDW surface. 19

2.1 This illustrates, in two dimensions, the grid based spatial indexing method. The naive S-GB method would require a distance computation to every other atom in the system. By only considering atoms in cells intersecting the radius of influence, represented here in blue, it is possible to consider far fewer interactions. Although only atoms inside the circle in this illustration contribute to surface charge, it is necessary to compute the distance over all black points. 29

2.2 Time spent during energy calculations on different parts of the energy model. The left pie represents the non-cell based approach and the right pie the cell based approach, the charts are scaled relative the total cost of computing the energy. One can see that although some overhead is introduced in maintaining the hash structure this significantly reduces the total cost of the solvent term, and as the solvent is such a large contributor to the total, also the total cost of computing the energy. 30

2.3 Energy computations using a grid based method yields approximately a three times performance improvement, though in the case of some very small structures it is possible that the overhead introduced by maintaining the grid structure outweighs the improvement. 32

3.1 The sequences which would be evaluated during an alanine scan for Fc domain of a human IgG for streptococcal protein G. The residues identified here were taken from the AESDB. The native protein is represented in the top row [Sauer-Eriksson *et al.*, 1995; Thorn and Bogan, 2001]. 38

3.2 Computed versus experimental $\Delta\Delta G$ binding for 8 alanine mutations in the Barstar-Barnase binding pair. Crystal structure used for computations was 1BRS. Specific amino acids mutated were residues 27, 54, 58, 59, 60, 73, 87, and 102, all of chain A. Experimental binding affinity taken from [Thorn and Bogan, 2001]. 41

3.3 Crystal, colored by atom, and predicted, magenta, side chain conformations for barnase, asparagine 58 of 1BRS. The predicted and crystal conformations are almost identical, differing by only 0.121 angstroms, or less than the resolution of the crystal structure. 43

3.4 Crystal, colored by element, and predicted, gray, side chain conformations of glutamic acid 73 of barnase, chain A of PDBid 1BRS. The two conformations differ by 0.993 angstrom RMSD, which is generally considered a successful sidechain prediction, though towards the upper range of a successful prediction. 43

3.5 Computed versus experimental $\Delta\Delta G$ binding for 6 alanine mutations in the Barstar-Barnase binding pair. Crystal structure used for computations was 1BRS [Buckle *et al.*, 1994]. Specific amino acids mutated were residues 29, 35, 39, 42, 74, and 78, all of chain D. Experimental binding affinity taken from [Thorn and Bogan, 2001]. 44

3.6 Distribution of 6 mutated residues, shown in magenta, on the interface surface of barstar, 1BRS chain D. Five of the six residues are less than 0.4 angstroms RMSD to the crystal structure. The only exception is, glutamic acid 80, shown in the upper left of this figure, and also [**?**]. 45

3.7 Crystal, colored by atom, and predicted, magenta, side chain conformations for barstar, chain D of PDBid 1BRS. The distance to the crystal structure is only 0.098 angstroms, or nearly identical. 46

3.8 Glutamic acid 80 is the only residue on chain D, barstar, of the barnase-barstar complex which was not predicted within 0.4 angstroms of the crystal coordinates during the mutation scanning experiments. The difference between these two conformations is 1.804 angstroms, which while sometimes considered a "successful" prediction, is not sufficiently close to generate the same interactions, making it difficult to accurately predict binding affinities. 46

3.9 Computed versus experimental $\Delta\Delta G$ binding for 10 alanine mutations in the anti-hen-egg-white lysozyme antibody (D1.3) anti-idiotopic antibody (E5.2) complex. Crystal structure used for computations was 1DVF [Braden *et al.*, 1996]. Specific amino acids mutated were residues 30, 32, 52, 54, 56, 58, 98, 99, 100, and 101, all of chain A. Experimental binding affinity taken from [Thorn and Bogan, 2001]. 48

3.10 An unsuccessful sidechain prediction in the antibody antigen complex of PDBid 1DVF. The predicted conformation of this aspartic acid, B:100, differs from the native state by 2.577 angstroms. 48

3.11 Two neighboring successful predictions, shown in magenta, in the same antibody antigen complex. Threonine 30, left, is predicted almost identically to the native structure, at 0.056 angstroms from the crystal coordinates. Tyrosine 32, right, is predicted at 0.412 angstroms RMSD. Predicted structures are shown in magenta in both cases. 49

3.12 Computed versus experimental $\Delta\Delta G$ binding for 8 alanine mutations in binding pair. Crystal structure used for computations was 1FCC. Specific amino acids mutated were residues 25, 27, 28, 31, 35, 40, 42, and 43, all of chain A. Experimental binding affinity taken from [Thorn and Bogan, 2001]. 51

3.13 Three clustered hot spot residues in another antibody antigen complex, PDBid 1FCC, the C2 fragment of streptococcal protein G in complex with the Fc domain of human IgG. The predicted conformations for glutamic acid 27, lysine 31 and tryptophan 43 are depicted in magenta. All three of these side chains are in good agreement with the crystal structure, with side chain RMSD's of 0.403, 0.594 and 0.371 respectively. These residues are shown in greater detail in other figures, glutamic acid 27 in figure 3.14, lysine 31 in figure 3.15, and tryptophan 43 in figure 3.16. Tryptophan 43 is interesting in that despite being critical to protein-protein binding it is largely buried in a pocket defined by the neighboring protein structure. This interaction is examined in figure 3.17. 52

3.14 Predicted, magenta, and crystal conformations, colored by element, for glutamic acid 27. The side chain RMSD of this prediction is 0.403 angstroms. This side chain is in close proximity to a number of other hot spot residues on the protein protein interface and is shown in context in figure 3.13. . . . 53

3.15 The predicted side chain conformation during the course of mutation scanning experiments, shown in magenta, compared to the native conformation, shown colored by atom for lysine 31. The RMSD of this prediction is 0.594 angstroms. 53

3.16 Native, colored by element, and predicted, magenta side chain conformation for tryptophan 43 of 1FCC. The root mean square distance of the predicted conformation to the native is 0.371 angstroms. The effect of the local protein structure on the conformation of this residue is examined in figure 3.17. . . 54

3.17 The pocket of tryptophan 43 of 1FCC. Because of conformation of the neighboring protein structure this residue has very little conformational freedom, and any prediction which successfully locates the sidechain in the pocket will be reasonably close to the native state. The conformation predicted in these experiments was very similar, 0.371 angstroms, and is depicted superimposed with the native in figure 3.16. 54

4.1 The structure of cytochrome P450, taken from PDBid 1JFB, shown in cartoon representation. The bonded heme group, shown as ball and stick model, is visible in the center. The brown iron atom is chelated by four deep blue nitrogen atoms. 58
4.2 An overview of the entire IDSite procedure. The dotted lines represent abbreviated versions of the full procedure. Receiver operating characteristic graphs for the full version, and these abbreviated versions, are presented in 4.15. Series colors on ROC graphs correspond to arrow colors here. 60
4.3 The bounding box used by Glide in order to generate the inital set of docked poses. The docking procedure also requires at least one hydrogen bond donor be found within 4 angstroms of the centroid of Glu216, Asp301, and Ser304 is also shown. The sphere representing this constraint is also shown. 62
4.4 The constraints applied to sp^2 atoms during the constrained minimization and first minimization Monte Carlo sampling stage. The spring constant of the bond constraint (red arrow) is 100 kcal/mol/angstrom^2, and that of the angle constraint is 25 kcal/mol/degree^2. The oxygen atom depicted in this figure is a "dummy" atom and does not interact with any other atoms in the structure except through the constraint. 64
4.5 The constraints applied to sp^3 atoms during the constrained minimization and first minimization Monte Carlo sampling stage. The spring constant of the bond constraint (red arrow) is 100 kcal/mol/angstrom^2, and that of the angle constraint is 25 kcal/mol/degree^2. The oxygen atom depicted in this figure is a "dummy" atom and does not interact with any other atoms in the structure except through the constraint. 66
4.6 The constraints applied to sp^2 atoms during the constrained minimization and second minimization Monte Carlo sampling stage. 66
4.7 The constraints applied to sp^3 atoms during the constrained minimization and second minimization Monte Carlo sampling stage. 67
4.8 The constraints applied to the salt bridge region of CYP2D6 during the *first* minimization Monte Carlo sampling stage. 67

4.9 The constraints applied to the salt bridge region of CYP2D6 during the *second* minimization Monte Carlo sampling stage. 68
4.10 An outline of the Monte Carlo minimization refinement stages in PLOP. . . 68
4.11 The linear relationship between the calculated intrinsic reactivity of the methoxy radical complex and that of the heme complex. Adapted from [Li *et al.*, 2011b] with minor correction. In the original manuscript the slope of the regression was reported as 1.117 and that number was used throughout. This difference should not significantly affect the physical IDSite classifier results, and does not affect the results of the fit model. In the rest of this text the value from the original publication of 1.117 will be used. 72
4.12 A comparison of the performance of IDSite with a variety of other methods of predicting P450 sites of metabolism. IDSite obtains the best performance, followed by a quantitative structure-activity relationship based method [Sheridan *et al.*, 2007]. Adapted from [Sheridan *et al.*, 2007]. . . . 76
4.13 Physical and fitted IDSite predictions of sites of metabolism on the training set. 79
4.13 (continued) 80
4.14 Physical and fitted IDSite predictions of sites of metabolism on the test set. 81
4.15 The effect of additional sampling on prediction of site of metabolism by P450. The light blue series describes only performing the initial Glide docking stage followed by minimization. The green series is obtained by using the set of structures obtained in the first minimization Monte Carlo sampling stage. The red series is obtained by screening the structures obtained in the first sampling stage, and minimizing these structures using the constraints specified in Figures 4.6 and 4.7. The blue series makes use of the entire IDSite procedure. The color scheme of these series corresponds to the colors of edges in Figure 4.2. 82

5.1 An overview of the regression testing procedure. 86

5.2 The fraction of minimized structures found within a given RMSD to native. The newer energy models, optimized variable dielectric (OVD or VSGB2.0) and variable dielectric surface generalized Born (VSGB) perform better than the original surface generalized Born model, however there is not significant differentiation between the two. 89

List of Tables

2.1 The specific timings for a series of energy computations presented in Figure 2.3. These represent a "best case" scenario, as the majority of time in these experiments is spent computing the solvent contribution. 33

3.1 Calculated and experimental $\Delta\Delta G$ for mutating given residues of barnase (chain A of structure 1BRS) to alanine. Experimental values taken from [Thorn and Bogan, 2001]. 42

3.2 RMSD of mutated side chains in barnase, in a barnase-barstar complex (chain A of PDBid 1BRS), during the mutation scanning experiments. 42

3.3 Calculated and experimental $\Delta\Delta G$ for mutating given residues of barstar (chain D of structure 1BRS) to alanine. Experimental values taken from [Thorn and Bogan, 2001]. 44

3.4 RMSD of mutated side chains in barstar, in a barnase-barstar complex (chain D of PDBid 1BRS), during the mutation scanning experiments. 45

3.5 Calculated and experimental $\Delta\Delta G$ for mutating given residues of anti-idiotopic antibody (chain A of structure 1BRS) to alanine. Experimental values taken from [Thorn and Bogan, 2001]. 47

3.6 RMSD of mutated side chains in 1DVF, anti-hen-egg-white lysozyme antibody (D1.3) complexed with an anti-idiotopic antibody (E5.2), during the mutation scanning experiments. 47

3.7 Calculated and experimental $\Delta\Delta G$ for mutating given residues of Fc domain of human IgG (chain A of structure 1FCC) to alanine. Experimental values taken from [Thorn and Bogan, 2001]. 50

3.8 RMSD of mutated side chains in 1FCC, C2 fragment of streptococcal protein G in complex with the Fc domain of human IgG, during the mutation scanning experiments. 51

4.1 The number of residues sampled as well as the number of structures advanced to the next stage from each of the sampling stages. Also, the relative probabilities of selecting each of the different sampling steps during a Monte Carlo minimization sampling stage. 63

4.2 DFT calculated values for internal reactivity of various compounds with either methoxy radical (compound I) or heme system. Correlation between these values is illustrated in Figure 4.11. 74

4.3 Results of physical and fitted IDSite on training set of 36 compounds. . . . 77

4.4 Results of physical and fitted IDSite on a test set of 20 compounds. Note that for the physical model there is no training performed so results in the text are presented in a unified fashion for the training and test set. 78

Chapter 1

Introduction

1.1 Drug Development

1.1.1 Costs of Drug Development

The process of bringing new drugs to market is a long and expensive affair. Information about both the costs and time necessary to bring a drug to clinical trials are less available than statistics for drug molecules reaching clinical trials. There is much debate over the average cost and time investment needed to develop a new drug. At the least it is necessary to identify a possible target molecule, find a small molecule with promising binding characteristics to that target which is additionally not toxic nor a strong binder to the wide variety of other proteins necessary for regular cellular function. These small molecules are then varied to maximize binding affinity to the target molecule, while attempting to simultaneously minimize cross reactivity. Finally, after this process, these drug compounds are rigorously tested through clinical trials. The final costs necessary for this process range from 400 million per new chemical entity to as much as 2 billion. Estimates for the time required also vary significantly, but many estimates place the time required at around 10 years from target identification to an approved drug. One of the largest factors affecting the average cost of each new drug compound is the low success rate for compounds which have been under active research for a number of years. Effectively screening these compounds earlier in the pipeline has the potential to significantly decrease the average cost of each

new drug molecule. [Adams and Brantner, 2006]

The average cost of identifying a new drug molecule and gaining approval for that molecule is actually growing at a rate greater than inflation. The number of new drugs introduced during the period 2005-2010 was actually 50% fewer than the number introduced during the time frame from 2000-2005. New drug compounds have been shown to have important impacts on both longevity and quality of life. In fact, during the 14 year period from 1986 to 2000, 40% of the two year increase in life expectancy can be accounted for by the effect of new drugs introduced during that period. [Paul *et al.*, 2010]

The expected period of time that a candidate drug compound will spend in clinical trials is approximately nine to fourteen years[DiMasi *et al.*, 2003; Paul *et al.*, 2010]. During the period from 1981 to 1990 the rate of approval of potential drugs decreased, as did that of self-originated drugs, or those drugs which were originally identified by a pharmaceutical research company. Of potential drug compounds which reach clinical trials, only 10% will finally be approved as new drugs [DiMasi, 2001; Paul *et al.*, 2010]. Of potential drug compounds entering clinical trials that fail to be approved as new drugs, approximately two thirds will be abandoned or fail during phase II clinical trials, which test the efficacy of a drug. This is generally viewed as a failure to find a small molecule with sufficiently high binding affinity to the target protein. Thirty percent of potential drug compounds entering clinical trials will fail in stage I, either because they are poorly tolerated, toxic to humans or cause side effects. Each of these is a potential indicator of cross reactivity with proteins other than the target molecule.

Computationally screening these compounds earlier in the process has the potential of reducing the attrition rate at this point in the process. Additionally, increasing the affinity for the target itself can allow for lower dosages which can increase survival through phase II clinical studies. Finally, approximately 20% of potential drug compounds entering clinical trials will fail in stage III. These drugs fail for a variety of reasons, though ineffectiveness is frequently cited as a reason. All told efficacy accounts for 37.6% of all drugs that are abandoned after reaching clinical trials, making it the single largest contributing factor to the failure of these compounds to eventually receive approval as new drugs. Other factors include saftey, and economics. [DiMasi, 2001]

For new chemical entities introduced in the 1990's the cost of research and development increasing at a rate 7.4% above inflation. Rates for the 2000's are not yet available or are only now becoming available due to the long lead time between introduction of a new chemical entity and that new chemical entity becoming an approved drug. During the period from 1985 to 2000 the rate of spending on research and development increased at approximately twice the rate of introduction of new chemical entities. Although the largest factors in determining this cost are the costs during clinical trials significant amounts are also spent earlier in the drug discovery pipeline, such as target identification, lead identification, and lead optimization. Improved computational techniques are generally viewed as possible means of decreasing costs or times associated with the earlier steps in the process. However, by increasing the fraction of leads which survive the screening process techniques which help identify and optimize lead molecules can have a very large effect on the cost of each new molecular entity. Clinical trials consist of six sometimes overlapping stages, denoted 0 to V, though stages I to III are where the majority of drug molecules are abandoned. Of the candidate compounds which enter clinical trials only approximately 20% will finally be approved as drugs. [DiMasi *et al.*, 2003]

Since 1950, the number of new chemical entities introduced per billion dollars has decreased by 50% every 9 years. Possible problems cited as contributing to this decrease in efficiency include:

1. the ready availability of high quality and effective generic drugs as treatment options for many diseases,

2. decreased risk tolerance among regulatory agencies,

3. increased spending and personnel without understanding underlying relationships between spending and personnel and discovery of new compounds, and the long period of time between beginning research on a drug target and finally gaining approval for a new drug compound, and

4. systematic overestimation of the efficacy of high throughput screening techniques relative more classical techniques such as clinical science, and animal screening [Scannell *et al.*, 2012].

The high failure rates during clinical trials have been identified as one of the most critical factors in determining the overall costs of drug development. [Bleicher *et al.*, 2003]

1.1.2 Computer Assisted Drug Design

The ultimate goal of computer assisted drug design is to improve rational drug design by exploiting the continuously increasing processing power available both in high performance super computers, but also in single workstations. Seeking to supplement the ability of a researcher either by allowing examination of a large number of possible interactions quickly or providing some insight that might be much more difficult to obtain through biochemical experiments, both in terms of time and expense. Different classes of programs have been developed to help solve each of the distinct steps in the pre-clinical stages of drug development, namely:

1. Hit Identification the process of screening a large small molecule database (up to one million or more small molecules) database to identify small molecules which bind a given target protein, or hits. These hits are usually small molecules with a target binding affinity on the order of micromolar.

2. Hit to lead optimization - the process of modifying these "hit" molecules either by substitution or addition of chemical moieties or mixing and matching substructures between given hits, to produce compounds with higher binding affinities than the initial hit compounds. Hit to lead optimization seeks to improve the micromolar binding affinity of hit compounds to nanomolar affinity or better.

3. Lead Optimization the final step of modifying lead compounds to increase "druglikeness" to ensure that the molecule is sufficiently soluble, well tolerated, and does not disrupt regular cellular function.

1.1.2.1 Hit Identification

The earliest form of hit identification experiments were animal screens, where mutant animals were studied to find the specific gene or protein causing a specific phenotype. This type of experiment relies on careful genetic controls and breeding, but also some

element of luck in observing a relevant phenotype in the first place. "Brute force" animal screens have since been improved with extensive mutation libraries and exhaustive non-lethal mutation libraries for organisms such as yeast and *Escherichia coli*. Even so, these screens are slow, often taking three years or longer, and error prone, as performing a large number of repetitive experiments causes even the most fastidious of scientists to lose focus. High-throughput screening seeks to supplement the human factor with robots, which are capable of performing similar experiments with greater speed and fewer errors. With the help of this automation it is possible to test the interactions of as many as 100 million different reactions per day [Agresti *et al.*, 2010]. Though the high initial cost of high-throughput screening equipment as well as the cost of the small molecule libraries necessary for screening are often prohibitive even to large research institutions. In order to make this sort of experiment available to a larger number of institutions some research institutions have instituted means of sharing this equipment, through high-throughput screening as a service type arrangements [HTSRC, 2004; MSSR, 2006].

The direct computational equivalent to high-throughput screening is virtual screening, where a library of small molecules is computational "docked" into the active site of the target protein, and some scoring metric is used to identify possible binders. In this sort of computational screen, he problem of the cost of small molecule libraries is essentially a solved problem in virtual screening as there are readily available libraries of drug-like small molecules for use in virtual screening programs. For example, the ZINC database provides a library of over seven-hundred thousand commercially available small molecules in a number of different file formats for use in virtual screening [Irwin and Shoichet, 2005]. Another possibility for hit identification *in silico* is through fragment assembly methods.

The first published study using computational docking was published in 1982 by Irwin Kuntz describing a program which would later go on to become the well known DOCK program [Kuntz *et al.*, 1982]. Generally docking consists of a method of quickly screening possible protein-small-molecule interaction conformations. An emphasis is placed on the computational cost of evaluating the energy function over accuracy, as the poses generated by this step are usually fed into structural refinement programs for further sampling and

more accurate estimation of energies. For example in the original Kuntz study, the system only only had six degrees of freedom on which to sample, three translational and three rotational degrees of freedom for the ligand with the protein held fixed. Along with a hard sphere collision model this provided a sufficiently selective screen to identify the native binding geometry of the heme group to myoglobin as well as thyroid hormone analogs to prealbumin [Kuntz *et al.*, 1982].

The rate at which new structures are being deposited into the Protein Data Bank is increasing on an annual basis. But tools are necessary to draw meaningful insights from this data, hopefully leading to new drugs.

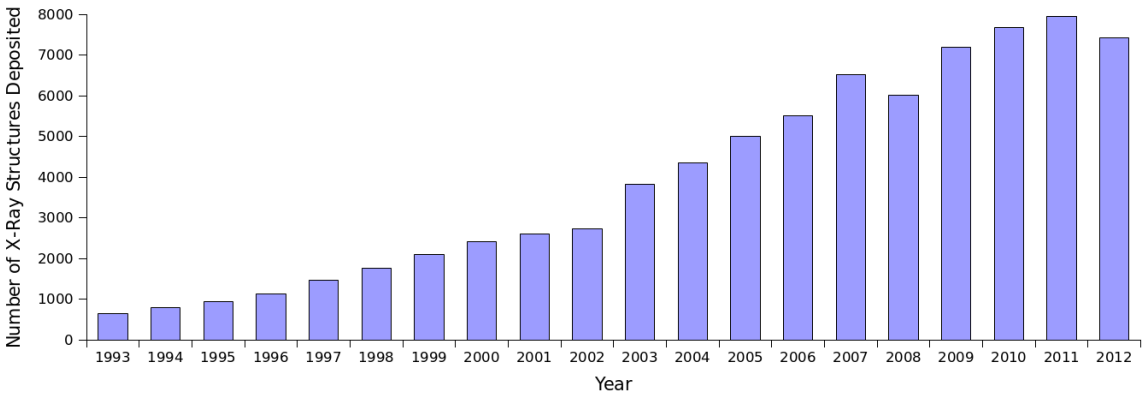


Figure 1.1: The rate at which new structures is deposited into the PDB over the last two decades. Due to a variety of improvements in the field of crystallography this rate has been steadily increasing.

For example a recent increase in the field of crystallography, is "crystal-less" crystallography in which small molecules are bound by a porous scaffold matrix. The regular structure of the matrix imparts a regular packing arrangement, necessary for interpreting diffraction patterns, onto the arrangement of small molecules. This has the potential to address one of the largest difficulties in obtaining quality structural data for proteins, which is that it is very difficult to purify and crystallize certain proteins [Inokuma *et al.*, 2013].

The number of target molecules of the set of all drugs currently available on the market consists of only about 500 proteins. The bottleneck in introduction of new chemical entities

is not virtual screening, but rather optimizing these hits into higher affinity leads and eventually balancing the requirements across all characteristics to produce a new drug [Bleicher *et al.*, 2003].

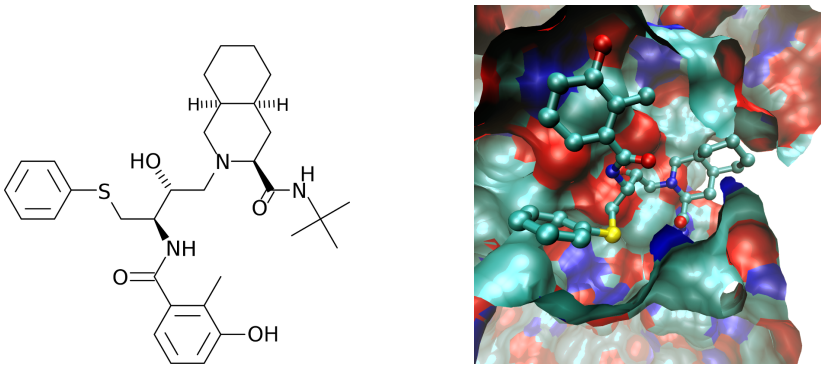
Of the total proteome only 30,000 are regulated by small molecule binding, making them reasonable targets of drug action. A large number of these possible drug targets are not implicated in any disease, due to this and a number of other factors, estimates of the total number of these proteins which are possible drug targets is much lower. Frequently cited numbers for the number of possible drug targets in humans are six-hundred to fifteen-hundred, still significantly higher than the total number of targets which are exploited by current drugs. The different families of cellular proteins are not equally likely to be targets of drugs. As of 2013 47% of current drug targets are enzymes, followed by 30% being GPCR's [Hopkins and Groom, 2002].

The consists of a number of characteristics which are generally true of drug like molecules:

1. Five or fewer hydrogen bond donors,
2. 500 Da or less total molecular mass
3. high liphophilicity
4. sum of nitrogen and oxygen atoms is not greater than 10 [Lipinski *et al.*, 1997]

Through understanding the protein-ligand conformation and specific contacts they were able to modify a known substrate There is an advantage to flexible substrates, which is that they can flex in order to create better contacts with the protein structure increasing binding affinity. This is especially important as the location of heavy atoms in the target protein is frequently only known to an accuracy of 0.4 angstroms. Further specific knowledge of the binding geometry between the initial lead compound and the target makes it possible to computationally screen possible chemical group substituents, to maximize binding affinity, increase solubility or bioavailability. One of the earliest examples of the successful application of structure based drug design is the carbonic anhydrase inhibitor dorzolamide, in which most of these ideas were applied to find a drug with very high binding affinity [Greer *et al.*, 1994].

Despite advantages in speed and cost due to limitations in accuracy computational screening has struggled to produce the same results as empirical screening. However, more recently virtual screening has succeeded in producing hit rates greater than those from empirical screening techniques. Virtual screening has been used to identified leads which were later developed into the human immunodeficiency virus (HIV) protease inhibitor Viracept, and the anti-influenza drug Relenza. A number of challenges which limit the utility of



(a) (b)

Figure 1.2: The HIV protease inhibitor, nelfinavir, marketed under the name Viracept was originally identified using a computational docking screen. It has a very high binding affinity for its target protein, 2 nM. Here it is shown crystallized with multidrug variant (ACT) (V82T/I84V) of HIV-1 protease, PDBID 3EL5. (b) generated with Visual Molecular Dynamics [Humphrey *et al.*, 1996] and [POVRAY 3.6, 2004].

docking programs have been identified

1. The number of possible small molecules is essentially unbounded, however only a very small fraction of these ligands are potentially drug compounds. Limiting sampling to this subspace is a challenging problem.

2. The number of conformations of ligand molecules rises exponentially with the number of internal degrees of freedom of the ligand. Sampling the huge conformational space of the ligand becomes a computationally difficult problem on its own.

3. The difficulty of accurately assessing or comparing the energy of different protein-

ligand complexes or conformations[Shoichet, 2004].

It has been found that introduced drugs are often very chemically similar the "hit" compounds from which they were derived [Proudfoot, 2002]. So in order to increase the diversity of drugs and find drugs which are able to treat new diseases, or diseases which have evolved resistance to current drugs it may be necessary to either increase the size of the screened database or increase the possible diversity which might be increase through the hit-to-lead step.

1.1.2.2 Hit-to-Lead Optimization

Hit compounds generally have a binding affinity for the target protein on the order of micromolar binding. The goals of hit-to-lead optimization are to further increase that affinity with the goal of eventually reaching binding affinities on the order of 10 nanomolar or better, find other molecules with similar chemical characteristics to increase the size and diversity of the set of lead compounds, and screening hit compounds for any obvious issues. At this stage for computational screening more accurate energy models are required than for the initial screen [Jorgensen, 2004; Gohlke and Klebe, 2002; Jorgensen, 2009].

Depending on the type of "hit" compounds identified in the initial screen, hits are either combined through molecular-growing and evolution techniques, or similar structures to the hit compounds can be sampled either by exploring the local chemical space or "mutation" of substituents. In either case, the potential lead compound is docked or grown in the known binding site.

A scoring function which is hopefully well correlated with the binding energy is then used to rank these possible compounds. Interestingly it is not necessarily the case that the scoring function is anchored in a physical force field, it is possible to use statistical or artificial intelligence approaches with success, so long as they are able to successfully solve the classification problem of distinguishing strong binders from weak binders. Docking as a means of converting hit compounds to lead compounds is very similar to docking as a means of hit generation, however in this case the small molecule library is much smaller and is generated to cover chemical space surrounding hit compounds. Additionally whereas for initial hit generation a coarse grained energy function might have been sufficient to

differentiate ligands which bind strongly from those which do not bind at all, to convert these "hits" to lead compounds it is necessary to use a more sensitive, and necessarily slower, energy model to accurately rank the binding affinity of different small molecules [Jorgensen, 2004; Gohlke and Klebe, 2002]. These energy models will be discussed briefly in 1.4.

A popular program for building, or mutating lead compounds is Biochemical and Organic Model Builder (BOMB) [Barreiro *et al.*, 2007]. BOMB can operated as either a hit identification program or as a hit to lead optimization method. Working to identify new compounds BOMB starts with a number of different small "core" scaffolds and attempts to increase binding affinity by adding or replace substituents with favorable interactions while avoiding steric clashes. BOMB has been successfully used to evolve a hit compound which showed no inhibition of HIV reverse transcriptase into a potent non-nucleoside RT inhibitor with nanomolar level binding [Barreiro *et al.*, 2007].

Whereas previously, lead compounds were evaluated almost exclusively on binding affinity to the target protein, more recently more weight is being placed on identifying hit compounds which satisfy other characterisitcs besides binding affinity [Bleicher *et al.*, 2003]. It is important to begin to consider other characterisitcs of the potential drugs earlier in the pre-clinical process, because later it is difficult to make changes which affect characterisitcs such as solubility without significantly altering the binding affinity of an already highly modified hit compound. As "lead" compounds are rarely very chemically distinct from the hits from which they were derived, and increasing binding affinity is actually sometimes an easier problem than addressing some of the other characteristics in the "rule of five" it is reasonable to begin by first trying to optimize hit compounds to satisfy some other criteria and postpone maximizing binding affinity [Proudfoot, 2002].

1.1.2.3 Lead Optimization

In lead optimization the compounds which have been identified by the earlier steps in the process are optimized to drug molecules. The largest differentiating factor between hit-to-lead optimization and lead optimization is the plausibility of the compound to act as a successful drug molecule. The goals of lead optimization overlap heavily with those of the

hit-to-lead stage. Although this can include increasing binding affinity to the target even further, usually the focus is on other characteristics including selectivity, ease of synthesis, pharmacokinetic properties and intellectual property concerns [Keserű and Makara, 2006]. Computational modelling can help not only identify hit compounds, and convert those initial hits into leads, but also to help estimate absorption, distribution, metabolism, elimination, toxicology, sometimes referred to as the ADME characteristics [Kerns and Di, 2008].

Computational models for ADME characteristics ususally use regression equations or neural networks to predict these characteristics [Jorgensen, 2004].

Up to one half of all drugs which do not survive clinical trials, fail to do so because of lack of efficacy, which is influenced both by binding, but also by the absorption characteristics of the molecule. The number of drugs which fail to make it through clinical trials due to toxcicity is similarly high, about 40% [Li, 2001]. Advancing a potential drug to clinical trials represents a very large financial investment, and effective computational screens of lead molecules at this point in the process can reduce the rate of failure in clinical trials, thereby having a very large impact on the final costs of new drugs brought to market.

1.2 Sampling Algorithms

1.2.1 Minimization

Minimization techniques seek to find the lowest energy conformation in a given potential well. Generally, they make no attempt to sample outside of that well, and therefore are frequently implemented as a final stage in sampling, in order to relieve any unfavorable interactions in proposed structures. There are a large number of different minimization techniques, and they will not be covered in any real depth here, please see the original papers for more details, or [Schlick, 2010] for a review. As the basic terms of the general molecular mechanics potential energy function are differentiable, and discounting for the moment the significant effects of solvent, it is possible to solve for the energy gradient, or force on every atom for a given conformation. A few minimizations methods include:

1. “Steepest descent”, conceptually the simplest minimization algorithm, in which the gradient is calculated at each step, and the size of the step is proportional to the

magnitude of the gradient [Levitt and Lifson, 1969; Bixon and Lifson, 1967].

2. "Newton" methods instead of approximating the gradient as a linear function in a small neighborhood, express the gradient, as a quadratic function. This has been shown to converge more quickly than steepest descent [Ponder and Richards, 1987]. Discrete Newton and Quasi-Newton methods use numeric estimation techniques instead of analytically solving for the gradient [Schlick, 2010].

3. "Truncated Newton" methods find an approximate solution to Newton's equations, forcing the residual to approach zero as the series converges [Dembo and Steihaug, 1983].

1.2.2 Monte-Carlo Sampling

Metropolis Monte-Carlo simulation was originally developed in the 1950's to provide rapid sampling of the solution space of many variable problems [Metropolis *et al.*, 1953; Hastings, 1970]. Monte-Carlo techniques generate a sequence of states from a distribution by proposing a new state based only on the current state If the ensemble average is the same as the sequence average, a Monte Carlo Markov chain can be used to estimate ensemble averages, this is known as *ergodicicy* [Schlick, 2010]. Another requirement is *detailed balance* that the probability of transition from a state X_i to a state X_{i+1} is the same as the probability of the reverse transition, i.e. X_{i+1} to X_i. By setting the probability of acceptance to

$$P(x \rightarrow x') = min\left(1, e^{-\frac{\Delta E}{k_B T}}\right) \tag{1.1}$$

these conditions are met.

In molecular mechanics, Metropolis Monte Carlo provides a very efficient means of sampling conformation space and a simple method of estimating the distribution of states. Modifications on this method, such as annealing, where the temperature is continuously decreased over the course of the simulation, or umbrella sampling, which attempts to achieve better sampling in cases where a potential energy barrier divides two or more states from each other [Torrie and Valleau, 1977]. While Monte Carlo sampling techniques are very fast to provide new states, the majority of these states reflect higher energy conformations.

Since it is of practical biological interest, Monte Carlo minimization has been developed to increase the rate at which minima are sampled [Li and Scheraga, 1987].

1.2.3 Analytic Loop Closure

Subsequences with regular secondary structures, α-helices and β-sheets are generally better conserved, and therefore likely to be well covered by simple homology models [Kolodny *et al.*, 2005]. The intervening "random coil" or loop regions often play a large role in determining protein specificity for a specific ligand as in antigen-antibody binding [Bajorath and Sheriff, 1996], small protein toxins to the receptors they target [Wu and Dean, 1996], or transcription factors to specific DNA sequences [Jones *et al.*, 1999].

Loop closure or prediction is a significant part of homology modeling, and building structures consistent with X-ray refraction data. Therefore in order to accurately predict three dimensional structure through homology models, infer the protein binding partners and function, or even build a three dimensional structure consistent with both X-ray data and physical constraints, accurately predicting these loop regions is critical [Fiser *et al.*, 2000].

The question is, given two fixed endpoints and a flexible loop, or actuator, find a conformation of the loop which connects the two endpoints. Because of the similarities that this problem solves with robotics a number of algorithms have been adapted from that field [Kolodny *et al.*, 2005]. The first of these is analytical loop closure, where a conformation which satisfies the closure criteria is solved for directly by solving a system of equations. Though this problem can be solved analytically for small loops [Wedemeyer and Scheraga, 1999; Go and Scheraga, 1970; Bruccoleri and Karplus, 1985; Palmer and Scheraga, 1991], the problem becomes more difficult as loop length grows and the number of degrees of freedom of the loop section increases. Additionally these closure constraints make sampling multiple different conformations more difficult [Cortés and Siméon, 2005], though it is possible to hierarchically solve sub-loops in order to generate conformations for possible loops [Wedemeyer and Scheraga, 1999].

1.2.4 Random Tweak

Random tweak, like CCD is a method of producing and sampling closed loop conformations. It begins in much the same was as CCD, by randomizing ϕ and ψ dihedral angles. Random tweak seeks to close the loop while retaining dihedral angles as close to the randomized starting structure as possible. By adjusting each dihedral only a small amount at a time and staying in the region where $sin(\Theta) \approx \Theta$ it is possible to formulate a set of linear equations to solve for a set of $\Delta\Theta_i$ which minimizes the distance between the crystal position of the atom to be closed and the random position. Because the assumption $sin(\Theta) \approx \Theta$, only holds for small Θ, the maximum change in angle is limited, 10 degrees in the original implementation. Because almost all structures predicted using the random tweak or cyclic coordinate descent produce closed loops, a much smaller fraction of time is spent sampling loops which do not satisfy the closure criteria, and these algorithms can be very efficient [Fine *et al.*, 1986; Shenkin *et al.*, 1987].

1.2.5 Cyclic Coordinate Descent

Another robotics algorithm which has been successfully applied to protein loop closure is Cyclic Coordinate Descent (CCD) [Canutescu and Dunbrack, 2003]. As the length of a flexible loop grows the number of degrees of freedom increases and the possible solution space grows exponentially. Cyclic coordinate descent seeks to close the loop by adjusting the degrees of freedom, in this case the ϕ and ψ dihedral angles, sequentially and possibly iterating over each degree of freedom multiple times until the loop is closed. This method is able to solve for conformations very quickly, and the likelyhood of closing a loop *increases* as the number of degrees of freedom of the system increases. In cyclic coordinate descent the ϕ and ψ angles of each loop backbone residue are first randomized. Then a loop dihedral is chosen at random, and varied to move the last atom of the loop as near as possible to its desired position. A new dihedral is chosen and optimized until the loop is closed. It is possible that this procedure does not converge to a closed state, however experiments have shown that this is very unlikely even for extended loops with few degrees of freedom, $< 2\%$ failure rate for 4 residue loops. Solving for the ideal dihedral angle at each step is a simple optimization problem making CCD a very fast algorithm [Wang and Chen, 1991;

Canutescu and Dunbrack, 2003]. In experiments CCD produces closed loop candidates in 1/6 the time of the random tweak method.

A variation on cyclic coordinate descent seeks to close the loop by not only requiring atom closure, but by requiring that the entire backbone of the closure residue is superimposed, within some geometric similarity tolerance, between the predicted and crystal structure. This constrainint ensures that the angles and dihedrals of the closure residue are reasonable[Canutescu and Dunbrack, 2003].

1.2.6 Rotamer Assembly

Rotamer assembly or systematic search shares some similarity with fragment buildup techniques in that it uses a rotamer library to assemble possible loops. This rotamer library represents the common backbone dihedrals for each amino acid. This method operates by dividing the loop into two pieces, usually in half, and considering all possible half loops which can be built using rotamer library [Moult and James, 1986]. For each side of the loop a "tree" is considered in both a physical sense, that the hemi-loop branches as it grows away from its anchor, and a decision tree sense, in that every residue represents a decision where a single rotamer is selected from the rotamer library. When the hemi-trees for each side of the gap are fully constructed some closure criteria is applied.

In the case of the original systematic search geometric agreement is required of the entire mid-residue [Moult and James, 1986], however a more lax criteria is applied in the case of the PLOP where only one atom is required to be approximately superimposed [Jacobson *et al.*, 2004]. By carefully pruning trees during the building process, and biasing the search towards occupied regions of ϕ-ψ space, systematic search can be quite efficient, spending little time sampling implausible regions of conformation space. Additionaly, by building residue pairs, using a smaller possibly restricting the rotamer library by building multiple residues at a time this sort of procedure has been used to build loops of 20+ residues [Zhao *et al.*, 2011].

1.3 Molecular Modeling

Molecular modeling seeks to gain new insights into the real world behavior of molecules by mimicking these molecules, usually using computer simulations. According to the theory of "minimal frustration" the protein native state is not only a low energy state, but is also stable [Bryngelson and Wolynes, 1987]. So the prediction of native or native-like conformations focuses on finding those conformations which have a low potential energy. As measuring the true potential energy of a system is very difficult or impossible computational models seek to reproduce the qualitative behavior of the energy surface. Quantum mechanics calculations are often viewed as the gold standard with respect to intramolecular energy calculations. However, despite the accuracy of quantum mechanics, its application to large systems such as proteins is currently limited due to the amount of time necessary to perform quantum mechanics calculations on a large number of atoms. Instead quantum mechanics calculations have been used to parameterize a majority of the most popular molecular mechanics force fields currently in use, including:

1. AMBER [Weiner *et al.*, 1984],
2. OPLS-AA [Kaminski *et al.*, 1994],
3. and CHARMM [MacKerell *et al.*,].

The earliest molecular mechanics force fields either modeled groups of atoms as a unit, hydrogens being grouped with their bound heavy atom [Jorgensen and Tirado-Rives, 1988], or even each residue as a unit [Lee *et al.*, 1999], both to reduce the number of parameters in the model and to increase the speed of computations. Although *ab initio* folding experiments are theoretically interesting, they are generally not practical both because of the difficulty in simulating such a large system for the time-frame necessary to observe behaviors like folding, and also because structural models for many proteins are available either directly as X-ray structures, or indirectly through homology.

Because of the evolutionary cost of mis-folded proteins, proteins have been selected to minimize mis-folding, making the general shape of the potential energy surface roughly funnel shaped with the native structure at the minimum [Leopold *et al.*, 1992]. Despite this

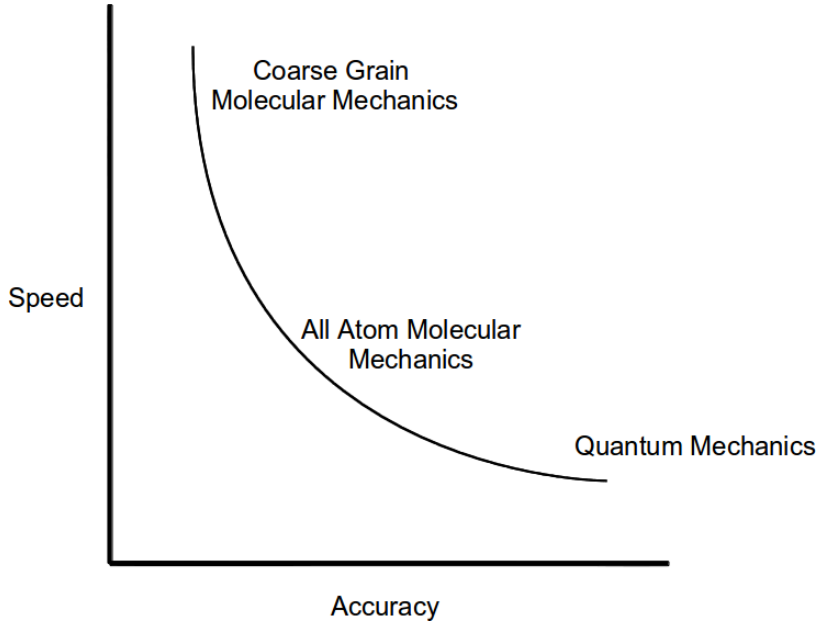
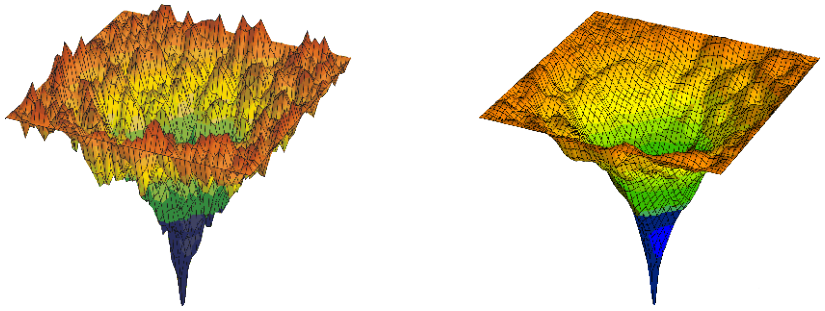


Figure 1.3: To an extent it is always possible to either increase accuracy or decrease running time, or the cost of an experiment. New scientific methods should allow one to increase accuracy while not spending additional time.

shape, the energy landscape of proteins is a very "jagged" surface with a large number of local minima [Tsai *et al.*, 1999].

Even the smallest enzyme contains 62 amino acids, and has thousands of degrees of freedom [Chen *et al.*, 1992], and larger enzymes are regularly more than 1000 amino acids. The number of degrees of freedom of these systems make any attempt to analytically solve for a global minimum energy conformation impossible, and require other methods of generating plausible conformations. In order to compensate for this a number of different sampling methods have been developed.



(a) (b)

Figure 1.4: Here energy is represented as a function of the two principal components of the protein conformation, in both cases the approximate funnel shape of the energy surface about the native conformation is very apparent. (a) shows an energy surface without any solvent effects. (b) represents a surface with the effect of hydration, though this surface appears smooth relative the dry surface in reality all energy landscapes of larger proteins contain many local minima. Figure from [Chaplin, 2013] used with authors permission.

1.4 Energy Functions

Some energy models do not seek to accurately rank potential conformations, fast "screening" functions attempt to quickly differentiate physically impossible conformations from plausible conformations without performing an expensive minimization or energy calculation step. Application of these screening functions has the potential to greatly reduce the number of potential conformations that must be scored using the full detail energy function, greatly decreasing the overall cost of conformation prediction. These screening criteria can be applied either during the sampling procedure, potentially eliminating sampling of a large area of excluded conformation space, or after sampling before a more expensive energy function is applied to rank conformations. Effective screening criteria have a large impact on the total performance of a structure prediction method.

One of the earliest screening criteria was the hard sphere overlap collision detection [Levinthal, 1966], and this method is consistently included in screening scriteria. Other screens include:

1. bounds on bond lengths and angles, as a single bond which deviates significantly from equilibrium can dominate the energy of a conformation,

2. limitations on ϕ-ψ space occuped by backbone dihedrals corresponding to the Ramachandran plot of the residue,

3. limiting side chain dihedrals to staggered conformations, which correspond to the low energy well of side chain dihedral space [Moult and James, 1986],

4. excluding structuctures which present excessive solvent accessible surface area, as this conflicts with the hydrophobic effect, which has a large effect on the conformation of the native state [Chothia and Janin, 1975]

5. limitations on the number of "dry" cavities, and the number of internal charged residues [Moult and James, 1986]

1.4.1 The General Form of the Energy Model

The general form of most molecular mechanics energy potentials is reasonably consistent, with bonds and angles being modeled as a spring, dihedrals as a Fourier series.

$$E\left(r^N\right) = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{dihedrals}} + E_{\text{nonbonded}} \tag{1.2}$$

$$E_{\text{bonds}} = \sum_{\text{bonds}} K_r(r - r_0)^2 \tag{1.3}$$

$$E_{\text{angles}} = \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 \tag{1.4}$$

$$E_{\text{dihedrals}} = \sum_{i=1\ldots4} \frac{V_i}{2}\left[1 + \cos\left(i * (\phi - \phi_0)\right)\right] \tag{1.5}$$

The non-bonded terms are modeled as a Columbic potential between any point charges and a Lennard-Jones or 6-12 potential between any non-bonded atoms. These non-bonded

atoms are phased in by a "fudge factor" for atoms in a 1-4 configuration.

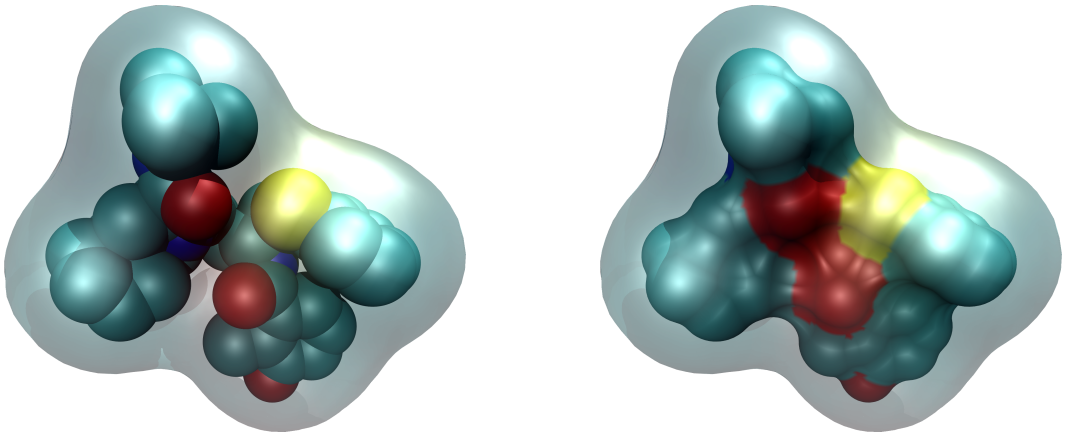
$$E_{\text{nonbonded}} = \sum_{i>j} f_{ij} \left(\frac{q_i q_j e^2}{r_{ij}} + 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right)$$

$$f_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ are separated by 2 or fewer bonds} \\ 0.5 & \text{if } i \text{ and } j \text{ are separated by 3 bonds} \\ 1.0 & \text{otherwise} \end{cases} \tag{1.6}$$

Where $\sigma_{ij} = \sqrt{\sigma_{ii}\sigma_{jj}}$ and $\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}}$ [Jorgensen *et al.*, 1996].

1.4.2 Molecular Surfaces

Central to the discussion of solvent is a discussion of how to formulate the surface of a protein. The most frequently used formulations of surface area include:



(a) (b)

Figure 1.5: (a) The Van der Waals surface of Nelfinavir, defined by the surface of the volume excluded by the VDW radii of the atoms in the structure. (b) The molecular surface, defined as the surface of the volume excluded from a probe of 1.4 angstroms (the radius of a water molecule). Both surfaces are enclosed by an approximate solvent accesible surface, which is defined as the surface traced by rolling a spherical probe over the VDW surface.

1. The Van der Waals (VDW) surface is the surface formed by the VDW radius of each molecule, though sometimes differing parameters are used for the VDW radii, and the ideal radii may even vary by application.

2. The solvent accessible surface, which is defined as the surface traced by the center of a spherical probe "rolled" over the VDW surface [Richards, 1977]. This idea is very closely related to the idea of the solvent excluded volume, or the shape of the solvent cavity enforced by the VDW surface of the molecule [Richmond, 1984].

3. The molecular surface, or Connolly surface, is composed of the VDW surface in areas where the spherical probe touches the VDW surface, in union with all points on the probe "between" two points on the VDW surface when the probe is contacting multiple atoms [Connolly, 1983], put another way the surface of the volume which intersects no possible probe location.

Frequently these surfaces are approximated numerically, using the Shrake-Rupley algorithm [Shrake and Rupley, 1973], by considering a spherical mesh about every atom and including only points which satisfy the definition of the surface, or using these points to interpolate a surface.

The significance of surface area is determined by physical constraints but can be well illustrated by a number of observations about proteins. First the ratio of total area of a theoretical unfolded, i.e. linearly arranged, protein to its length is almost among proteins, only varying by ~3% between different proteins.

1.4.3 Solvent Models

Beyond covalent terms and electrostatic effects, solvation effects have a very large effect on protein structure, and the interactions between proteins and small molecules. Therefore it is critical to accurately model the effect of the solvent on the molecule. While explicitly modeling each water molecule and sampling over possible conformations is the most realistic possible model, doing so requires calculating both a large number of solute-solute interactions as well as sampling extensively different solvent configurations. In this case it is likely that more time will be spent determining the behavior of the solvent than that of the solute. Because of these complexities even with efficient methods of sampling explicit solvent models, these simulations are too expensive to use on systems the size of proteins [Figueirido *et al.*, 1997; Zhang *et al.*, 2001].

Therefore there is significant interest in continuum models which accurately describe the mean force of water, without requiring additional sampling or interactions as in explicit models [Zhang *et al.*, 2001; Still *et al.*, 1990; Qiu *et al.*, 1997]. These methods have the potential to be three orders of magnitude, or even more, faster than explicit solvent experiments, and a number of different methods have been shown to accurately describe solvent effects [Zhang *et al.*, 2001].

The total free energy of solvation can be separated into polar and non-polar components, which correspond to the work done inserting the uncharged solute molecule, or protein, into the solvent and then building the charges to their native values [Roux and Simonson, 1999].

$$E_{solvent} = \Delta W_{\mathrm{non-polar}} + \Delta W_{electrostatic} \tag{1.7}$$

According to scaled particle theory the non-polar work done by inserting a sphere into a solvent can be approximated if the raidus of the sphere is neither too large nor to small as

$$\Delta W_{np}(s) = \gamma SA(X) \tag{1.8}$$

where γ is the surface tension of the solvent and the surface area corresponds most closely to the Connolly surface.

The electrostatic contribution to the solvent energy is the work necessary to add a charge to a hard sphere atom already in the solvent. The charge density in the solvent can be given by the Poisson-Boltzmann equation

$$\nabla \cdot [\epsilon(r)\nabla\psi(r)] = -4\pi\rho_u(r) \tag{1.9}$$

Though it is possible to solve this at every step of a simulation, it becomes rather expensive, therefore faster approximations are sought [Nicholls and Honig, 1991]. The total work can be approximated by the Born model

$$\Delta W_{electrostatic} = \frac{Q^2}{2R}\left(\frac{1}{\epsilon_v} - 1\right) \tag{1.10}$$

However, this assumes that the induced charge in the solvent is entirely concentrated on the surface of the ion, which is impossible, therefore R, or the Born α radius becomes a fit parameter, representing the effective radius of a charged sphere in the solvent.

The electrostatic solvation contribution can also be expressed as

$$\Delta W_{electrostatic} = \frac{1}{2}\sum_{i,j} q_i q_j f(x_i, x_j) = \frac{1}{2}\sum_{i} q_i^2 f(x_i, x_i) + \frac{1}{2}\sum_{i \neq j} q_i q_j f(x_i, x_j) \tag{1.11}$$

where f is a weighting function for the interaction between charges q_i and q_j. Historically there are a variety of methods of approximating this weighting function, however one of the most popular is the generalized Born (GB) [Still *et al.*, 1990]. In the generalized Born approch

$$f(x_i, x_j) = \sqrt{d(x_i, x_j) + R_i R_j e^{\frac{-d(x_i, x_j)^2}{4R_i R_j}}} \tag{1.12}$$

and one of the limiting factors to accuracy becomes obtaining proper estimates of the effecitve radii, since charges are not uniformly exposed to the solvent [Schaefer and Karplus, 1996].

Another approach is to estimate both the nonpolar and electrostatic contributions to solvation as proportional to the surface area, with different proportionality constants for different atoms.

$$\Delta W = \sum_{atoms} \gamma_{atom} SA(atom) \tag{1.13}$$

Although this method is very inexpensive to compute, it can be somewhat difficult to solve for the force on an atom, due to the way the surface changes as atoms move [Roux and Simonson, 1999].

Chapter 2

A Cell Based Method for Evaluating Implicit Solvation Effects

One can see in the Generalized Born model (1.12), computing the electrostatic contribution to solvation requires computing the distance between every two atoms.

2.1 Introduction

Computational protein structure prediction and related areas of research such as target screening and lead optimization continue to be areas of active research in both pure chemistry and pharmaceutical applications [Jorgensen, 2009]. These methods range from identifying leads using chemical similarity metrics, to artificial intelligence methods such as neural networks and support vector machines, to structural based methods [Geppert *et al.*, 2010]. In the recent past, structural methods have contributed to the identification of bioactive drug compounds [Corsino *et al.*, 2009], making computational protein structure prediction highly important to the medical field, and because of its pharmaceutical applications, economically relevant as well.

There are over 81,000 X-ray structures presently in the PDB, more than 8,500 of which have been added in the last 12 months, and the rate at which new structures are determined

by X-ray crystallography continues to accelerate [Berman *et al.*, 2007]. The chemical space of small molecules, i.e. potential drug compounds, is essentially unlimited, and at the least too large to effectively screen using conventional experimental methods [Jorgensen, 2009]. Furthermore, computational loop prediction experiments are predicting longer loops, increasingly relying on the output of initial predictions as the input to a later "fixed stage" refinement step that re-predicts some central region of the same loop. This has been shown to increase accuracy [Jacobson *et al.*, 2004] at the cost of an increased number of experiments and corresponding increase in computational cost. Taken together, these factors necessitate the development of more accurate and efficient methods of generating and evaluating protein-protein and protein-small molecule conformations and interactions.

The Protein Local Optimization Program (PLOP), originally developed by Friesner and co-workers, is a popular program used to predict, sample, and evaluate protein conformations [Jacobson *et al.*, 2002a; Jacobson *et al.*, 2002b; Jacobson *et al.*, 2004]. PLOP makes use of the OPLS-AA energy model, an atomic detail force field optimized for organic, including protein, interactions [Jorgensen *et al.*, 1996]. In addition to the terms defined by the OPLS-AA model, it has been shown that solvent effects can have a large contribution to prediction accuracy. The solvent contribution to an energy model can be evaluated either by explicitly modeling and sampling solvent molecules (usually water) or by treating the solvent as a continuous medium, i.e. implicit solvation. PLOP, like many other molecular mechanics programs, makes use of an implicit solvent model. This is largely because explicitly modeling solvent molecules, while possibly very accurate, requires extensive, and therefore very time consuming, sampling of a large number of small molecules [Zhang *et al.*, 2001]. Further, energy errors using explicit solvent models can be due to either an unrealistic force field parameters or insufficient sampling of solvent molecule conformations [Zhou, 2003]. Continuum solvation methods, or implicit solvent methods, attempt to address these issues by removing the dependence on sampling of solvent molecules and introducing approximations that reduce the cost of calculating the solvent contribution to energy [Roux and Simonson, 1999].

Among implicit solvent methods, the surface area based generalized Born (S-GB) model is one of the most popular and has been shown to produce results in good agreement with

experimental data [Zhang *et al.*, 2001; Gallicchio *et al.*, 2002]. The surface area based generalized Born implicit solvent model provides an approximate solution to the Poisson–Boltzmann equation based on a surface integral [Ghosh *et al.*, 1998]. However, in a naive implementation of this model, the electrostatic contribution is a sum over every charge-charge pair in the system, in this case the protein and its crystal copies.

$$U = \sum_{\text{charges}} U_{self}(q_k, r_k) + \sum_{\text{charges, i} \neq \text{j}} U_{pair}(q_i, q_j, r_i, r_j) \tag{2.1}$$

Electrostatic contribution to solvation free energy, decomposed into self, and pair terms. The time complexity of evaluating the pair term is quadratic in the number of charges in the system [Ghosh *et al.*, 1998]. This means that calculating the electrostatic solvation contribution for a single atom requires a linear search over all other charges in the system, which for large systems becomes the bottleneck of the computation. A common optimization is to assume that the electrostatic contribution to the solvation term for point charges separated by a distance greater than a defined cutoff distance is negligible [Gallicchio and Levy, 2004]. However, making this assumption does not improve the underlying quadratic time complexity of evaluating the solvation term, as it is necessary to compute the inter-charge distance for every charge pair in the system, before possibly excluding the interaction. In the implementation of S-GB in the Protein Local Optimization Program, computing the solvation term of a large structure is the rate limiting step of energy calculations, accounting for over 80% of the total time spent computing the energy (see figure 2.2). Therefore, less expensive methods of evaluating the implicit solvent contribution to the energy of a system can allow for increased sampling with the same available resources, thus improving efficiency of computational modeling.

We present an application of a geometric hashing method, grid based spatial indexing, to implicit solvent calculations in PLOP. The hashing method proceeds by dividing space into cubical regions, or cells, and distributing atoms into those cells, while maintaining a list of the contents of each. Retrieval of atoms within a cell can then be performed in constant time, and retrieval of a superset of atoms contained within a region can be performed in time proportional to the number of cells intersecting the region. This efficient geometric lookup allows one to replace a loop over all atoms inside the structure with a loop over

only the atoms contained in cells intersecting the sphere with radius corresponding to the distance cutoff of the force in question. While maintaining a list of atoms for each of these grid boxes introduces some overhead when updating atomic coordinates during a simulation, updating atomic coordinate is still a constant time operation. Thus, the benefits outweigh the costs, especially for large systems. As long as the cell size is bounded below, the number of cells necessary to consider when evaluating a fixed distance interaction is constant. Physical limitations provide an upper limit to atom density. Constant time retrieval of the contents of each hash cell, along with upper bounds on the number of atoms per cell and the necessary number of cells to consider for each charge, guarantee constant time lookup of all atoms within a given sphere. This reduces the time complexity of evaluating the electrostatic contribution from $O(n^2)$ to $O(n)$. We show that an implementation of this hashing method can reproduce results obtained with a non-hash based implementation while providing significant performance improvements.

2.2 Methods

2.2.1 Energy Model

The energy model used in side chain prediction experiments was the optimized variable dielectric model (OVD), sometimes referenced as the variable dielectric surface generalized Born 2.0 model (VSGB2.0). This energy model is based on the OPLS-AA energy model, which in terms gets most of its covalent parameters from the AMBER force field [Jorgensen *et al.*, 1996]. The solvation term used is a surface area based generalized Born formulation, where the internal dielectrics of charged amino acids have been optimized over a set of 2239 single side chain predictions and 100 loop predictions of 11 to 13 residue loops. In addition to the covalent terms from the OPLS-AA model the current energy model also includes terms to describe $\pi - \pi$ stacking, hydrogen bonding, and a parametrized hydrophobic term for the non-polar free energy of solvation [Li *et al.*, 2011a]. For the electrostatic contribution to solvation free energy two different molecular surfaces are maintained at different resolutions. A surface mesh is constructed for each atom using the generalized spiral points method [Rakhmanov *et al.*, 1994; Saff and Kuijlaars, 1997; Zhou, 1995], the number of points for

each sphere is 10 for the low resolution surface and 330 for the high resolution surface. An atomic based distance cutoff is used when evaluating the electrostatic contribution to the solvation free energy. Inside a distance of $\sqrt{50}$ angstroms the high resolution surface is used, between this distance and 20 angstroms the lower resolution surface is used, and the contribution of atoms outside this distance is assumed to be negligable. The same set of cutoffs is used in both the current implementation in PLOP and the new cell based method described here.

2.2.2 Data Sets

The data set used for energy calculation experiments consisted of large protein structures, containing neither DNA, RNA, or modified residues with molecular masses between 100 and 150 kDa. All samples had resolutions better than 2 angstroms. The data set was filtered at 30% sequence identity, and from this, 20 structures were selected at random, though two were later excluded due to technical reasons.

The structures used in side chain prediction experiments consisted of high resolution enzyme structures. Structures without an enzyme classification were excluded, as were structures with X-ray resolution less than 1.5 angstroms, or those with modified protein residues. This resolution requirement was imposed to make high resolution side chain prediction comparisons more meaningful. All structures had a molecular mass between 11 kDa and 110 kDa. Structures containing DNA, RNA or modified side chain residues were excluded, as were those that had unreasonable steric clashes either within the canonical structure or with crystal neighbors. Structures containing certain small molecules without energy parameterizations currently defined within PLOP were also excluded.

2.2.3 Structure Preparation

In preparing the crystal structures for energy calculation and side chain prediction experiments, the first step was to add the crystal copies of the protein of interest. PLOP completes this step for all space groups according to the crystal symmetry identified in the PDB file. Before modeling a structure using an all atom force field, it is also necessary to add hydrogens and any missing heavy atoms. When possible, PLOP uses the positions of

bonded heavy atoms to build missing atoms into the structure. However, adding hydrogens, especially for titratable residues, is a more complicated problem. To address this, PLOP uses the independent cluster decomposition algorithm (ICDA) to determine the protonation states of any titratable residues, as well as the positions of polar hydrogens [Li *et al.*, 2007]. Generally speaking, this proceeds by dividing titratable and polar residues into independent groups using a distance cutoff, and optimizing each group independently. Structures with unreasonable steric clashes with crystal neighbors were removed from the data set on the basis that such structures are physically unlikely.

2.2.4 Grid-Based Spatial Indexing

Grid-based spatial indexing is a well known algorithm in computer science, especially computer graphics, that allows for efficient lookup based on geometric criteria and also provides fast collision detection [Bentley and Friedman, 1979]. Critical to the present application, it allows constant time retrieval of a superset of atoms guaranteed to contain all atoms within a given euclidean distance. In our implementation, the bounding box of the protein and its symmetric copies is subdivided along each of the orthogonal axes to form grid boxes or cells. A simple convention for handling atoms that are positioned along cell boundaries guarantees that each atom is hashed to a unique cell. For a single dimension, d, the cell index, or hash, of a point p is

$$i_d(p) = int\left(N * \frac{p_d - min_d(P)}{max_d(P) - min_d(P)}\right) \tag{2.2}$$

where $min_d(P)$ and $max_d(P)$ are the minimum and maximum coordinates in dimension d over the set of points P, N is the number of cells in dimension d and p_d is the coordinate of p in d. Following the same procedure in each dimension gives a unique cell location for every atom. In this way, at the beginning of the simulation, each atom is assigned to a specific cell, or grid box. A list of the atoms in each grid box is then maintained over the course of the simulation. When computing the electrostatic contribution to solvent free energy of an atom, a, it is only necessary to loop over the atoms contained in boxes that intersect the sphere corresponding to the distance cutoff around atom a. Beyond that cutoff, charge effects are considered to be negligible [Gallicchio and Levy, 2004].

In the present implementation, cell size is at first set to 2.745 angstroms, and the number of cells in a given dimension depends on the "length" of the system in that dimension. If the number of cells that this would require is unmanageably large, cells are then grown simultaneously in all dimensions such that the cell size is 1/250th of the longest dimension of the structure.

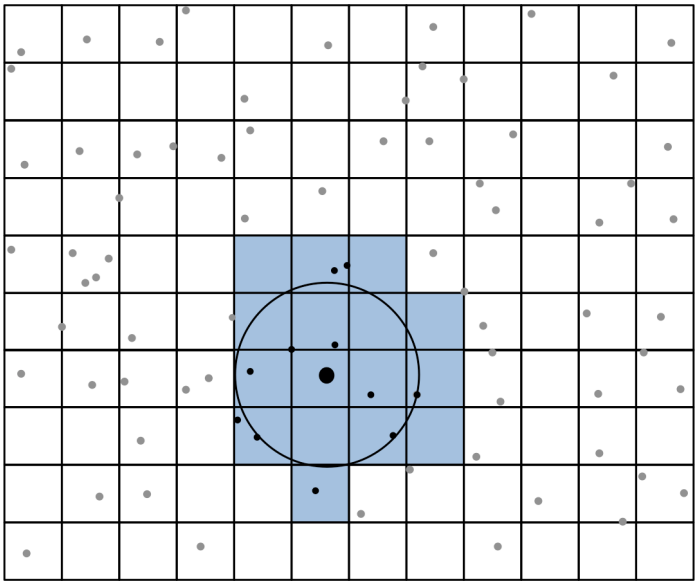


Figure 2.1: This illustrates, in two dimensions, the grid based spatial indexing method. The naive S-GB method would require a distance computation to every other atom in the system. By only considering atoms in cells intersecting the radius of influence, represented here in blue, it is possible to consider far fewer interactions. Although only atoms inside the circle in this illustration contribute to surface charge, it is necessary to compute the distance over all black points.

2.2.5 Experiments

For side chain prediction, the specific side chains used were those which had at least 30% solvent accessible surface area when evaluated in the absence of other chains or crystal neighbors. Glycine and proline residues were also excluded, as they do not have free side chains. Residues missing heavy atoms in the crystal structure were predicted; however,

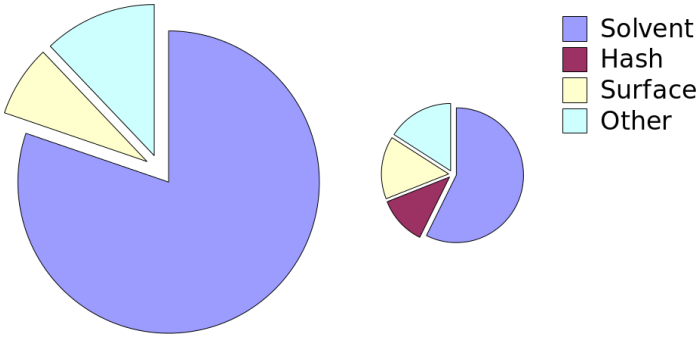


Figure 2.2: Time spent during energy calculations on different parts of the energy model. The left pie represents the non-cell based approach and the right pie the cell based approach, the charts are scaled relative the total cost of computing the energy. One can see that although some overhead is introduced in maintaining the hash structure this significantly reduces the total cost of the solvent term, and as the solvent is such a large contributor to the total, also the total cost of computing the energy.

RMSD was not measured for these residues because there is no experimental data. Side chain prediction experiments were performed as described in [Jacobson *et al.*, 2002a].

The experiment in this case consisted of multiple energy calculations, using the modified version of the OPLS-AA force field described in Li *et. al.* 2011 [Li *et al.*, 2011a]. In the control experiments, the same method for evaluating the implicit solvent term was used as in previous works.

2.3 Results

Qualititave Measures of Prediction Quality

For side chain prediction experiments, 85.2% of side chain prediction conformations (9406 of 11030 total) predicted with the new cell based solvation model are within 0.2 angstrom heavy atom RMSD of the prediction using the naive implementation. In other metrics, the quality of prediction is comparable between the two solvent models. Median side chain heavy atom RMSD is 0.567 and 0.558 angstroms for the cell based method and the non-cell based method, respectively. Average RMSD to the crystal structure is similarly close,

1.11 angstroms for both methods, with 79.9% of side chain predictions within 2 angstroms RMSD of the native using the cell based model and 79.4% within two angstroms using the naive approach. Of side chains which are predicted differently by the two implementations there is no correlation between solvation model and prediction quality. The distribution of side chain predictions with respect to RMSD to native is also indistinguishable between the two methods of computing the solvation term.

Data for energy calculations is not presented here because it is identical in every case. This is expected, given that the two models represent two methods of computing the same quantity. Thus, on the whole, prediction accuracy of the hash based model is comparable with the old implementation.

Performance Improvement

The principal goal of the hash based approach is to improve the performance of the implicit solvent models. Thus, the key metric of performance improvement is the speedup over the previous implementation. Energy computations were found to be from 1.6 to 2.5 times as fast, and the trend indicates that even larger improvements would be obtained in larger system. This sort of experiment represents a "best case" for the expected performance increase of a hash based solvent, as they represent a minimum amount of time spent on other parts of the experiment. Implicit solvent calculations, and energy calculations in general, compose a smaller fraction of time in simultaneous side chain prediction therefore the observed performance improvement is less than that of energy calculations. The observed performance increase in this sort of experiment is still on approximately 20%.

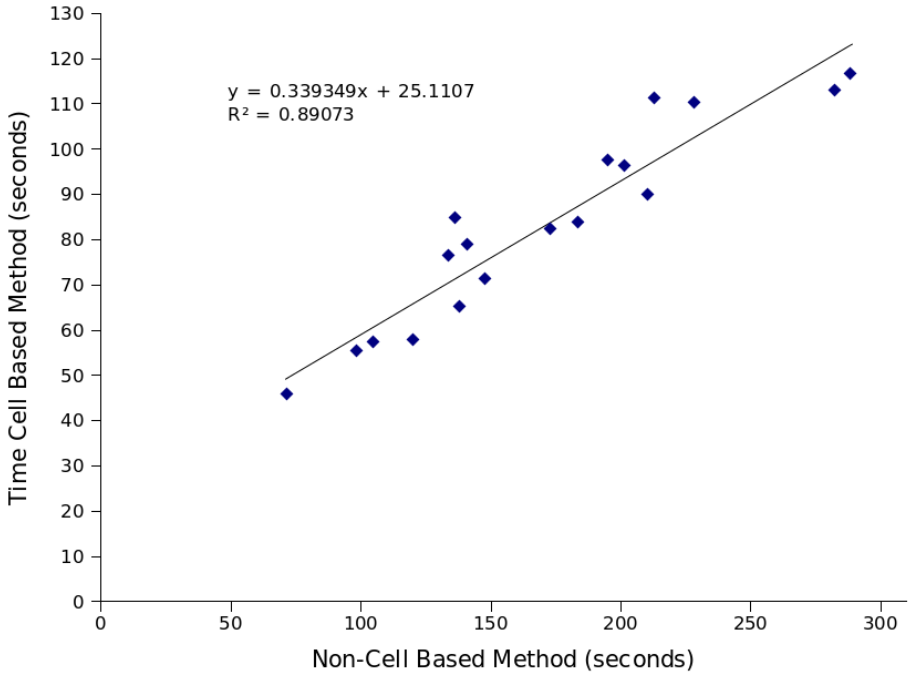


Figure 2.3: Energy computations using a grid based method yields approximately a three times performance improvement, though in the case of some very small structures it is possible that the overhead introduced by maintaining the grid structure outweighs the improvement.

PDB id	Naive Method	Cell Based Method
1F5Z	201.75	96.37
1H2V	98.31	55.32
1HRD	228.38	110.12
1M1Z	147.83	71.28
1M9X	282.35	113.04
1O60	172.82	82.34
1R0V	210.35	90.0
1XMP	213.25	111.25
2E3Z	141.11	78.79
2H6U	138.25	65.07
2OU1	104.93	57.24
2XI9	71.68	45.73
3AMD	288.58	116.52
3DEL	133.62	76.48
3E1E	183.8	83.8
3FGN	120.38	57.75
3HHP	195.25	97.58
4GVR	136.35	84.78

Table 2.1: The specific timings for a series of energy computations presented in Figure 2.3. These represent a "best case" scenario, as the majority of time in these experiments is spent computing the solvent contribution.

2.4 Discussion

We have developed an application of a classic computer science grid based hashing algorithm to the implicit solvent model of PLOP. We demonstrate that this application does not affect accuracy of results compared with the previous implicit solvent model implementation in PLOP. Though in a small number of cases side chain conformations are predicted in widely

different conformations by the hash based method and the old implementation, the two methods are equally likely to predict the more native-like structure, so this variance can be attributed to noise. We have also found in other experiments that the final predicted structure of a minimization is very sensitive to both small changes in pre-minimization coordinates, of a magnitude far less than bond distances, and minimization parameters. It possible that these effects magnify small differences present early in the experiment, resulting in much larger differences between the final predicted structures. Finally, we present data showing that the reduced computational cost of evaluating the solvent contribution using this hash based approach dramatically reduces the total time spent evaluating the energy model by a factor of 1.6 to 2.5.

Note that any such geometric hashing will introduce some overhead for maintaining the data structure. As discussed by Bentley and Friedman [Bentley and Friedman, 1979], the total storage necessary for the hash structure and the time necessary to sort atoms into cells are both linear in the number of atoms, and placing or updating a single atom in the structure is a constant time operation. The improvement in retrieval using this structure dominates the cost of maintaining the structure, and the difference becomes more pronounced as system size grows. Bentley and Friedman also present a through review of the performance characteristics of a number of other geometric hashing techniques, though they compare the algorithms in a data agnostic means. Taking advantage of the characteristics of physical data, in this case atomic coordinates, has some effect of the relative advantages and disadvantages of specific hashing techniques. Specifically, the maximum number of atoms per cell is limited by physical constraints of atomic interactions. An octree is a similar, though hierarchical, hash structure used in computer graphics for fast location based retrieval. However, because the criteria for "collision" in this case is a fixed distance cutoff, and the data is roughly uniformly distributed it is efficient to used a fixed cell size [Turk, 1989].

Though the implicit solvent term was initially targeted, because it dominates the time spent in energy calculations, it is possible to apply this method to any pair-pair interaction. Though especially for shorter range interactions it might be beneficial to either maintain a higher resolution hash, or implement a hierarchical spatial hash, such as an octree. We

are also applying this geometric hashing technique to collision detection between simultaneous loop predictions. This will allow efficient screening of neighboring loop prediction candidates, which will be particularly useful in predicting structures with multiple nearby solvent exposed loops, such as G-protein coupled receptors.

Implicit solvation models offer a very tangible benefit over explicit solvent models, both in performance and experimental complexity. Though explicit solvation is sometimes viewed as a "gold standard", it has been shown that current implicit solvent models can, at least sometimes, reproduce predictions of explicit solvent models. However, development of implicit solvent models is important since improved performance compared with explicit solvent methods allows modeling of larger systems, longer timescales, or improved sampling. The complexity of the experiment is also reduced, using implicit solvent models, because results are not dependent on sampling of water conformations in addition to protein conformations. However, implicit solvent models can still be very computationally expensive. For instance, in the PLOP implementation of the OPLS-AA energy model with SG-B solvation term, evaluating the solvation term consumes up to 80% of the total time spent in energy calculations, dependent on size of the symmetric system. This is in large part due to the time complexity of evaluating the SG-B solvation term. Thus an algorithm that offers further reduction in experimental cost without a tradeoff in accuracy represents valuable progress in the development of implicit solvent models.

Because the size of protein systems are limited, at some level, by physical limits, the maximum system size expected to be encountered is limited. Therefore, although the new method reduces the time complexity of implicit solvent calculations, the maximum expected speedup is limited to about a factor of three in large systems. The actual speedup depends on both the system size, and the amount of time that a given experiment spends evaluating the solvent contribution. The speedup observed in side chain prediction experiments was much less, around 20%, though it is possible that applying a similar method to terms of the gradient during minimization would increase that amount. Nonetheless, even a 20% speedup represents a significant improvement, especially as structure prediction methods continue to depend on parallel prediction and reprediction of the same region as a method of structure refinement [Goldfeld *et al.*, 2013]. Although this algorithm improves on the

theoretical time complexity of the S-GB implicit solvent model, parameterization such as the number, and size of cells in the grid structure could have a significant effect on run time. Some effort was made towards choosing reasonable parameters, but they are likely not optimal. Hardware that is optimized for this sort of spatial indexing and collision detection, or proximity detection, exists in modern video cards, and along with general purpose programming for this sort of hardware it should be possible to further parallelize computation of implicit solvation effects for even greater performance improvements [Harris, 2008].

Chapter 3

Progress in Computational Mutation Scanning

3.1 Introduction

In order to determine which amino acids in a protein play the largest role in determining binding affinity it is convenient to compare the binding affinity of the native protein with that of a single residue mutant. If the change in residue does not affect the folding of the protein, which with the exceptions of mutations to glycine, proline, or depending on the local structure possibly large amino acids, is unlikely. Alanine is the most frequently occurring amino acid, appearing in both solvent exposed and buried positions [Chothia, 1976; Rose *et al.*, 1985], and is unlikely to disrupt the protein fold in the same way glycine or proline might [Klapper, 1977]. Additionally because it lacks a charge it does not interact electrostatically with the ligand. These reasons makes it an attractive choice as a "control" amino acid for mutation scanning experiments. Mutation scanning experiments, seek to identify residues which are have the largest contributions to binding affinity or "hot spot" residues. By identifying single residue mutants which have a significantly decreased binding affinity when mutated to alanine [Cunningham and Wells, 1989].

The immune system maturation response, selects antibodies which have a reasonable affinity for an antigen, and creates a large number of variants of these antibodies. The effect of this is that the body produces antibodies with increasing affinity for an antigen some

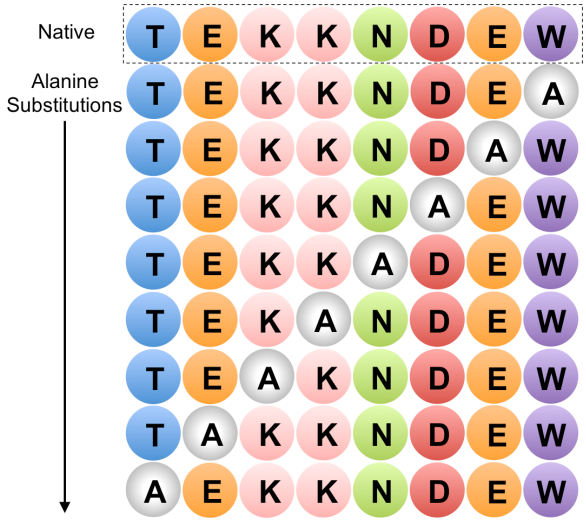


Figure 3.1: The sequences which would be evaluated during an alanine scan for Fc domain of a human IgG for streptococcal protein G. The residues identified here were taken from the AESDB. The native protein is represented in the top row [Sauer-Eriksson *et al.*, 1995; Thorn and Bogan, 2001].

time after the initial exposure [Griffiths *et al.*, 1984]. In vitro affinity maturation attempts to select molecules, frequently antibodies, with high affinity for some target molecule by creating a library of bacteria displaying variants of these antibodies on their cellular surface. The means of doing this is bacteriophage display, which provides a method of pairing the protein represented on a bacterias surface with the genetic material contained by that bacteria [Smith, 1985]. A bacteria is infected by a library of bacteriophages, containing a large number of variants of antibody of interest. The phage will cause the bacteria to display their specific variant of the antibody on the bacteria surface, allowing sorting of the bacteria according to their affinity of the antigen, though affinity column purification or similar technique. This step greatly enriches the fraction of antibody variants which bind the protein. It is then possible to allow the bacteria to reproduce, sometimes causing more mutations to increase diversity of the antibody library and perform this affinity purification step again. Sequential application of this affinity maturation makes it possible to identify a handful of antibody variants with high affinity from as many as 10^6 different variants [Gram *et al.*, 1992; Hawkins *et al.*, 1992]. However, the number of possible variants of

the complementarity determining region (CDR) of an antibody many orders of magnitude larger than this.

Computational mutation scanning attempts to replicate the same sort of experiment *in silico*. Making the same assumptions as above, namely that the backbone conformation is not altered by mutating a single residue to alanine, computational experiments attempt to identify hot spot residues by measuring the $\Delta\Delta G$ between the bound states of the native and mutated protein.

Mutated structures were generated by truncating sidechain at C_γ [Massova and Kollman, 1999]

Varying cutoffs for *hot spot* residues are used, usually from 1.0 kcal/mol [Kortemme and Baker, 2002] to 4.0 kcal/mol [Pons *et al.*, 1999].

Mutations at these *hot spot* residues tend to be strongly deleterious leading to above average conservation [Hu *et al.*, 2000; Lichtarge *et al.*, 1996].

3.1.1 Entropy-Enthalpy Compensation

Some computational alanine scanning experiments explicitly compute or approximate the entropic contribution to the change in the free energy of binding [Hao *et al.*, 2010; Guerois *et al.*, 2002]. However, other models have acheived good agreement with experimental data while assuming these effects are either accounted for by the correlation between entropy and enthalpy for small changes in protein structure [Sharp, 2001] or to be small relative the entropic changes [Kortemme *et al.*, 2004]. PLOP has not made use of entropic contributions to free energy and has in many cases acheived good agreement with experiment, so in these experiments it is assumed that contributions due to entropy are small relative entropic contributions.

3.2 Methods

3.2.1 General Mutation Screening

The generalized mutation screening method implemented in PLOP allows efficient evaluation of a large number of possible mutations. It accepts as input a set of possible mutations

for each residue, or a set of possible mutations for a set of residues. For instance tryptophan, tyrosine and arginine are overrepresented in hot spot resiudes [Hu *et al.*, 2000], so it may be desirable to consider all mutations in which a set of residues are either left at their native identity or replaced with one of these residues. If desired the user can also set bounds for the minimum and maximum number of simultaneous mutations allowed. The residues which will be mutated are referred to as *free* residues as the conformations of the other residues are held fixed throughout the entire process. While the residues are still in their native states the structure is subjected to some sort of sampling. This is done in order to prevent bias towards predicted states which will later be predicted in the same fashion. In the present implementation this consists of predicting the conformation of each free residue is minimizing those residues.

In this side chain sampling, for each free residue, the sidechain is initially replaced with a random conformation from a high resolution rotamer library screened for steric clashes with the static part of the protein. The free residues are then examined sequentially replacing each with the lowest energy conformation present in the rotamer library. This replacement process is continued until the termination condition is met, which is that two or fewer residues are replaced by lower energy conformations during the replacement stage. Five iterations of this procedure, from randomization to a static conformation, are performed and the most frequently selected conformation is chosen for each amino acid [Jacobson *et al.*, 2002b; Jacobson *et al.*, 2002a].

In the mutation stage, each free residue is first updated to its new chemical identity, possibly remaining in the native state, sidechains are replaced with the sidechain of the desired amino acid, with the corresponding updates to the bond, angle, torsion, and 1-4 interactions. The conformations of free residues are then re-predicted using the same side chain prediction algorithm described for the native conformation.

3.2.2 Alanine Scanning Experiments

Three protein complexes, 1FCC, 1BRS, and 1DVF, with both experimental data for binding affinity and crystal structures were identified using the ASEdb [Thorn and Bogan, 2001]. Protonation states and locations of polar hydrogens were assigned for all residues as in [Li *et*

al., 2007]. A crystal context was built for each structure using symmetry data determined by experiment. For each mutation represented in the alanine scan database single residue was mutated to alanine and this side chain prediction was repeated. The resulting structures were examined side chain conformation agreement with crystal structures and the change in binding free energy to native was recorded.

3.3 Results

A strong correlation between observed and expected $\Delta\Delta G$ was not found for any of the structures nor for the set of mutations taken as a whole. Despite this single side chain conformations were in very good agreement between predicted sidechain locations and crystal structures, suggesting that sufficient sampling was done in the side chain prediction step.

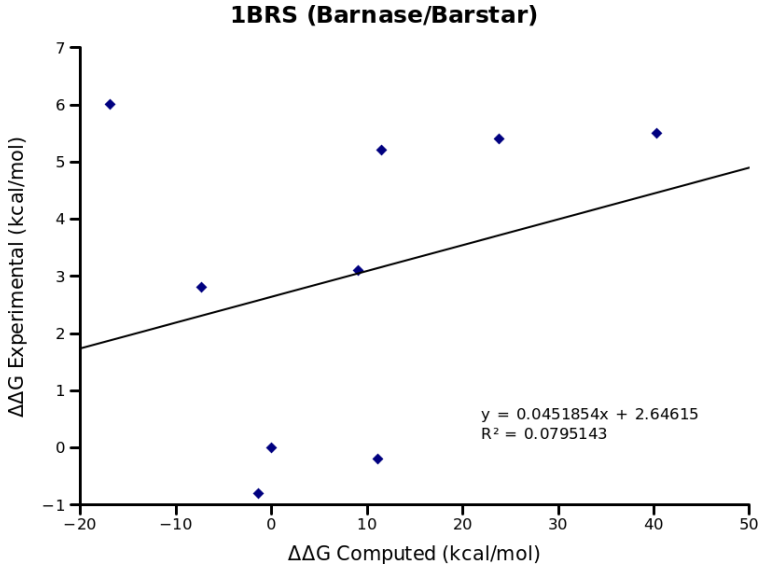


Figure 3.2: Computed versus experimental $\Delta\Delta G$ binding for 8 alanine mutations in the Barstar-Barnase binding pair. Crystal structure used for computations was 1BRS. Specific amino acids mutated were residues 27, 54, 58, 59, 60, 73, 87, and 102, all of chain A. Experimental binding affinity taken from [Thorn and Bogan, 2001].

Residue	$\Delta\Delta G$ calculated	$\Delta\Delta G$ experimental
native	0	0
27	23.82	5.4
54	-1.37	-0.8
58	9.09	3.1
59	11.58	5.2
60	11.15	-0.2
73	-7.28	2.8
87	40.32	5.5
102	-16.83	6

Table 3.1: Calculated and experimental $\Delta\Delta G$ for mutating given residues of barnase (chain A of structure 1BRS) to alanine. Experimental values taken from [Thorn and Bogan, 2001].

Residue	Amino Acid	RMSD
A:27	LYS	1.213
A:54	ASP	0.920
A:58	ASN	0.121
A:59	ARG	0.421
A:60	GLU	0.138
A:73	GLU	0.994
A:87	ARG	0.297
A:102	HIS	0.276

Table 3.2: RMSD of mutated side chains in barnase, in a barnase-barstar complex (chain A of PDBid 1BRS), during the mutation scanning experiments.

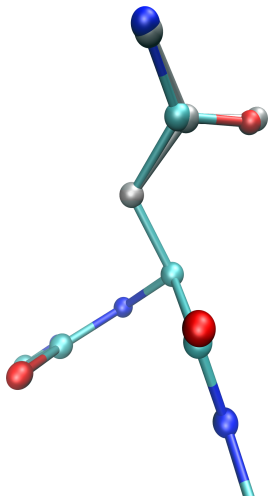


Figure 3.3: Crystal, colored by atom, and predicted, magenta, side chain conformations for barnase, asparagine 58 of 1BRS. The predicted and crystal conformations are almost identical, differing by only 0.121 angstroms, or less than the resolution of the crystal structure.

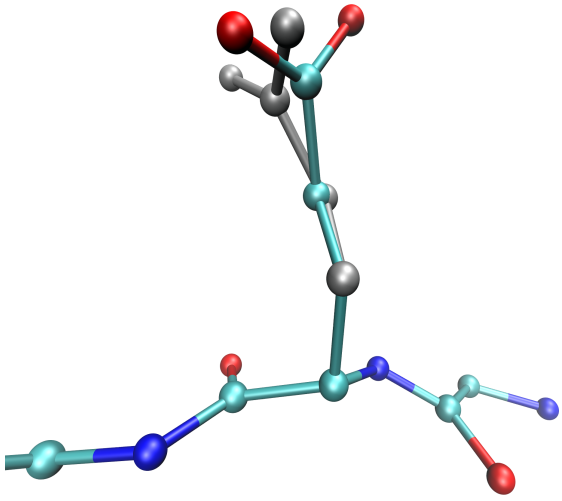


Figure 3.4: Crystal, colored by element, and predicted, gray, side chain conformations of glutamic acid 73 of barnase, chain A of PDBid 1BRS. The two conformations differ by 0.993 angstrom RMSD, which is generally considered a successful sidechain prediction, though towards the upper range of a successful prediction.

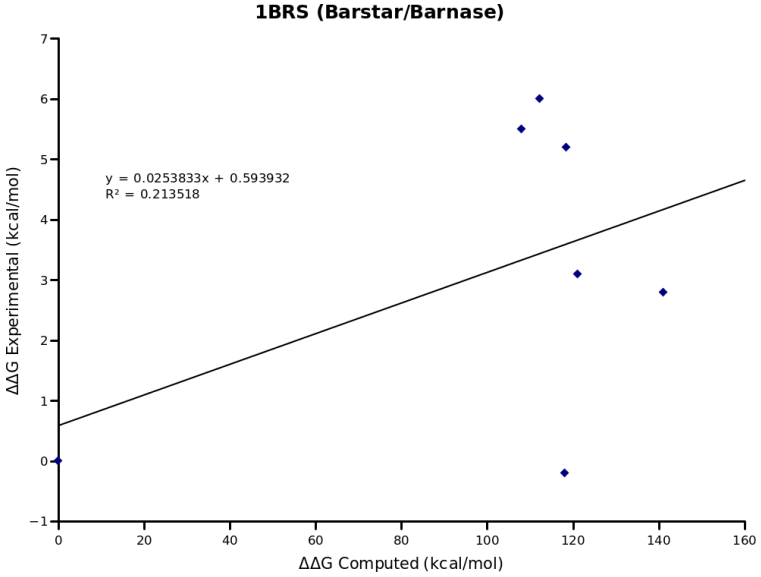


Figure 3.5: Computed versus experimental $\Delta\Delta G$ binding for 6 alanine mutations in the Barstar-Barnase binding pair. Crystal structure used for computations was 1BRS [Buckle *et al.*, 1994]. Specific amino acids mutated were residues 29, 35, 39, 42, 74, and 78, all of chain D. Experimental binding affinity taken from [Thorn and Bogan, 2001].

Residue	$\Delta\Delta G$ calculated	$\Delta\Delta G$ experimental
native	0	0
29	121.07	3.1
35	118.37	5.2
39	118.09	-0.2
42	141.09	2.8
74	107.92	5.5
78	112.14	6

Table 3.3: Calculated and experimental $\Delta\Delta G$ for mutating given residues of barstar (chain D of structure 1BRS) to alanine. Experimental values taken from [Thorn and Bogan, 2001].

Residue	Amino Acid	RMSD
D:29	TYR	0.121
D:35	ASP	0.098
D:39	ASP	0.335
D:42	THR	0.114
D:76	GLU	0.397
D:80	GLU	1.804

Table 3.4: RMSD of mutated side chains in barstar, in a barnase-barstar complex (chain D of PDBid 1BRS), during the mutation scanning experiments.

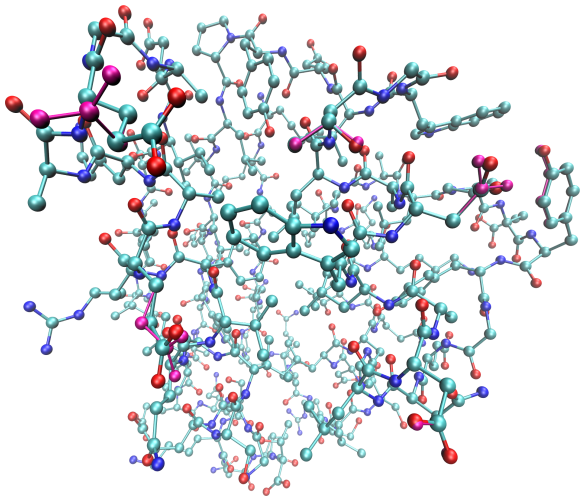


Figure 3.6: Distribution of 6 mutated residues, shown in magenta, on the interface surface of barstar, 1BRS chain D. Five of the six residues are less than 0.4 angstroms RMSD to the crystal structure. The only exception is, glutamic acid 80, shown in the upper left of this figure, and also [**?**].

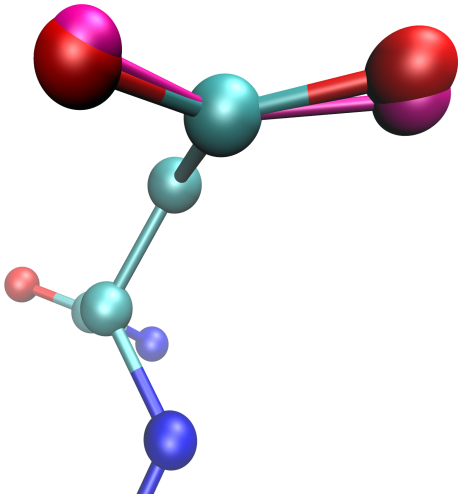


Figure 3.7: Crystal, colored by atom, and predicted, magenta, side chain conformations for barstar, chain D of PDBid 1BRS. The distance to the crystal structure is only 0.098 angstroms, or nearly identical.

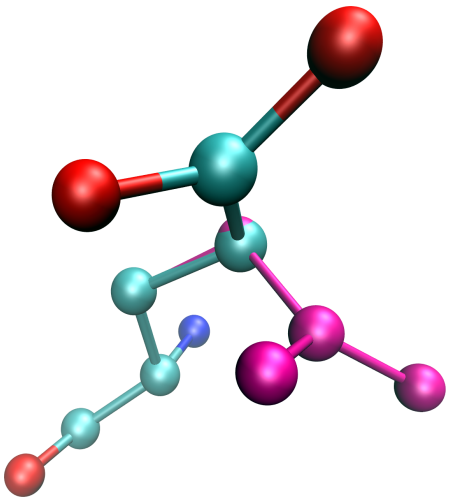


Figure 3.8: Glutamic acid 80 is the only residue on chain D, barstar, of the barnase-barstar complex which was not predicted within 0.4 angstroms of the crystal coordinates during the mutation scanning experiments. The difference between these two conformations is 1.804 angstroms, which while sometimes considered a "successful" prediction, is not sufficiently close to generate the same interactions, making it difficult to accurately predict binding affinities.

Residue	$\Delta\Delta G$ calculated	$\Delta\Delta G$ experimental
native	0	0
30	-42.93	0.9
32	-44.68	1.8
52	-42.6	4.2
54	-28.29	4.3
56	-37.5	1.2
58	-41.91	1.6
98	-34.38	4.2
99	-30.51	1.9
100	-60.9	2.8
101	-37.84	4

Table 3.5: Calculated and experimental $\Delta\Delta G$ for mutating given residues of anti-idiotopic antibody (chain A of structure 1BRS) to alanine. Experimental values taken from [Thorn and Bogan, 2001].

Residue	Amino Acid	RMSD
B:30	THR	0.056
B:32	TYR	0.412
B:52	TRP	0.391
B:54	ASP	0.442
B:56	ASN	0.238
B:58	ASP	0.159
B:98	GLU	0.182
B:99	ARG	1.137
B:100	ASP	2.577
B:101	TYR	0.340

Table 3.6: RMSD of mutated side chains in 1DVF, anti-hen-egg-white lysozyme antibody (D1.3) complexed with an anti-idiotopic antibody (E5.2), during the mutation scanning experiments.

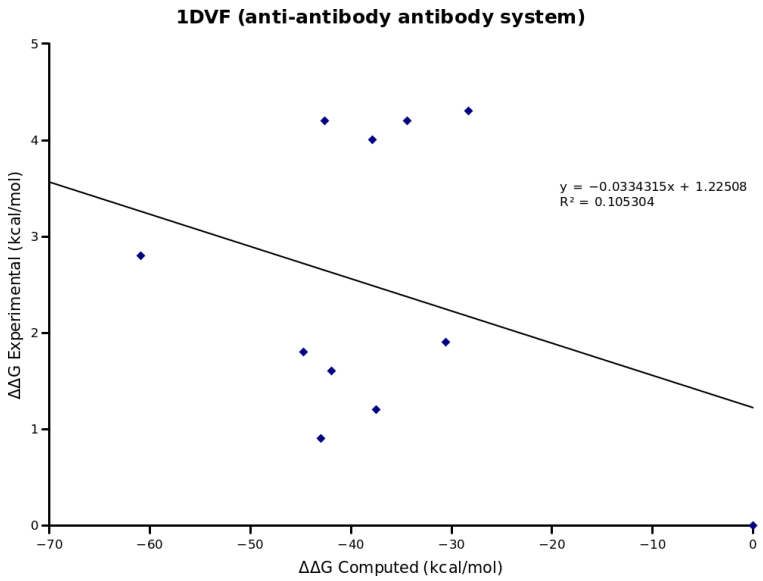


Figure 3.9: Computed versus experimental $\Delta\Delta G$ binding for 10 alanine mutations in the anti-hen-egg-white lysozyme antibody (D1.3) anti-idiotopic antibody (E5.2) complex. Crystal structure used for computations was 1DVF [Braden *et al.*, 1996]. Specific amino acids mutated were residues 30, 32, 52, 54, 56, 58, 98, 99, 100, and 101, all of chain A. Experimental binding affinity taken from [Thorn and Bogan, 2001].

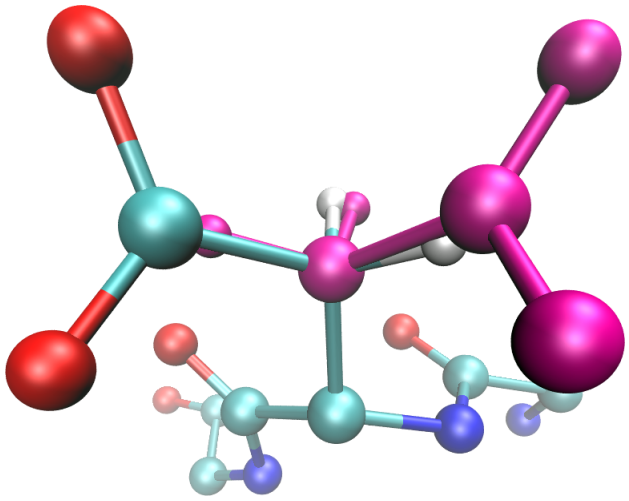


Figure 3.10: An unsuccessful sidechain prediction in the antibody antigen complex of PDBid 1DVF. The predicted conformation of this aspartic acid, B:100, differs from the native state by 2.577 angstroms.

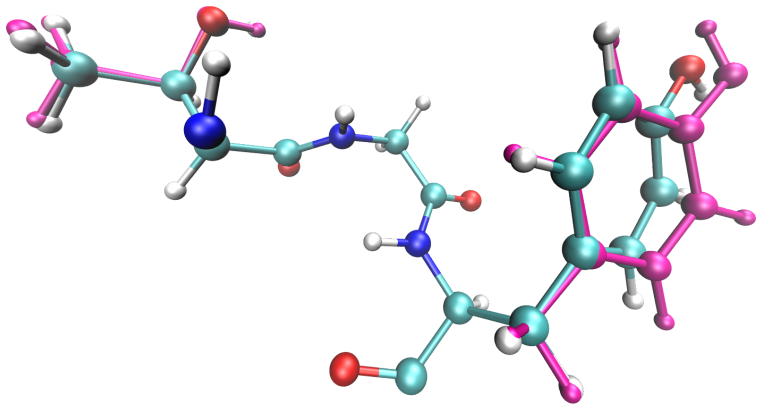


Figure 3.11: Two neighboring successful predictions, shown in magenta, in the same antibody antigen complex. Threonine 30, left, is predicted almost identically to the native structure, at 0.056 angstroms from the crystal coordinates. Tyrosine 32, right, is predicted at 0.412 angstroms RMSD. Predicted structures are shown in magenta in both cases.

Residue	$\Delta\Delta G$ calculated	$\Delta\Delta G$ experimental
native	-0.18	0
25	-4.55	0.24
27	19.8	4.9
28	26.37	1.3
31	18.82	3.5
35	-6.31	2.4
40	-8.51	0.3
42	-5.56	0.4
43	12.0	3.8

Table 3.7: Calculated and experimental $\Delta\Delta G$ for mutating given residues of Fc domain of human IgG (chain A of structure 1FCC) to alanine. Experimental values taken from [Thorn and Bogan, 2001].

Residue	Amino Acid	RMSD
C:25	THR	0.170
C:27	GLU	0.403
C:28	LYS	0.543
C:31	LYS	0.594
C:35	ASN	0.527
C:40	ASP	1.292
C:42	GLU	2.994
C:43	TRP	0.371

Table 3.8: RMSD of mutated side chains in 1FCC, C2 fragment of streptococcal protein G in complex with the Fc domain of human IgG, during the mutation scanning experiments.

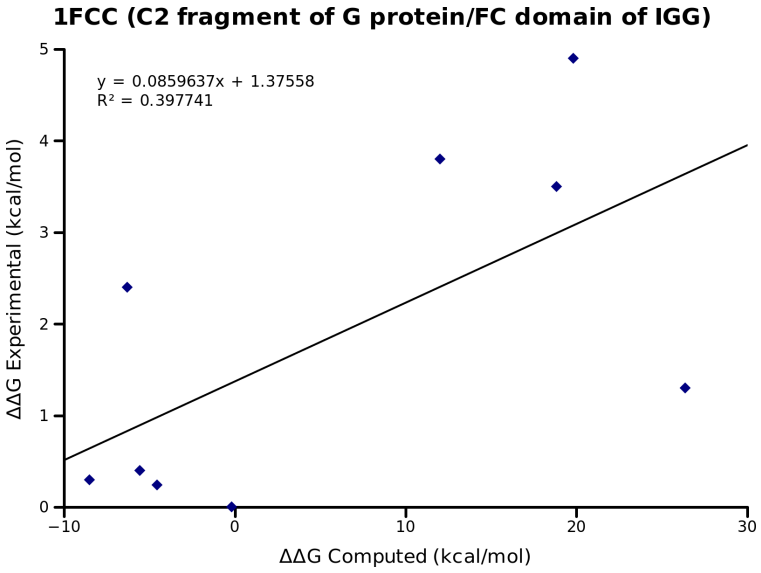


Figure 3.12: Computed versus experimental $\Delta\Delta G$ binding for 8 alanine mutations in binding pair. Crystal structure used for computations was 1FCC. Specific amino acids mutated were residues 25, 27, 28, 31, 35, 40, 42, and 43, all of chain A. Experimental binding affinity taken from [Thorn and Bogan, 2001].

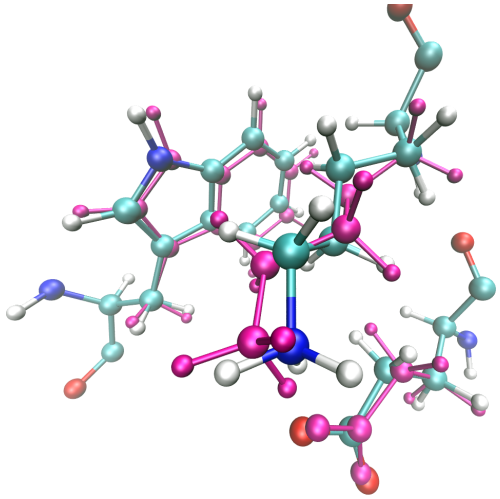


Figure 3.13: Three clustered hot spot residues in another antibody antigen complex, PDBid 1FCC, the C2 fragment of streptococcal protein G in complex with the Fc domain of human IgG. The predicted conformations for glutamic acid 27, lysine 31 and tryptophan 43 are depicted in magenta. All three of these side chains are in good agreement with the crystal structure, with side chain RMSD's of 0.403, 0.594 and 0.371 respectively. These residues are shown in greater detail in other figures, glutamic acid 27 in figure 3.14, lysine 31 in figure 3.15, and tryptophan 43 in figure 3.16. Tryptophan 43 is interesting in that despite being critical to protein-protein binding it is largely buried in a pocket defined by the neighboring protein structure. This interaction is examined in figure 3.17.

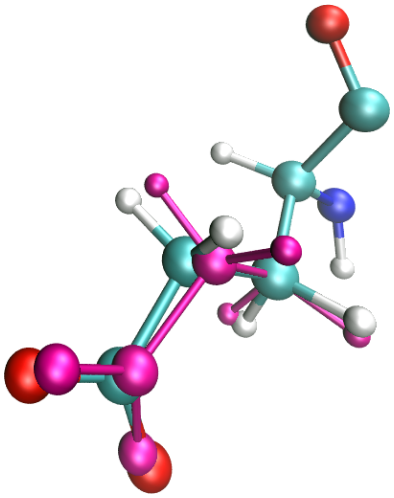


Figure 3.14: Predicted, magenta, and crystal conformations, colored by element, for glutamic acid 27. The side chain RMSD of this prediction is 0.403 angstroms. This side chain is in close proximity to a number of other hot spot residues on the protein protein interface and is shown in context in figure 3.13.

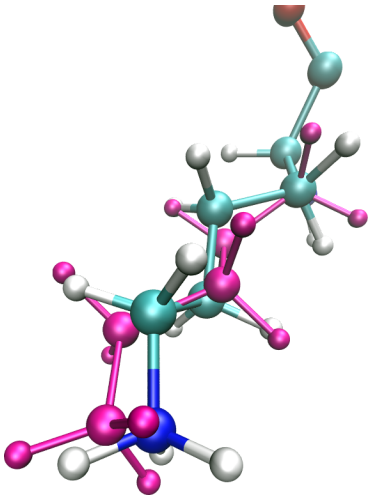


Figure 3.15: The predicted side chain conformation during the course of mutation scanning experiments, shown in magenta, compared to the native conformation, shown colored by atom for lysine 31. The RMSD of this prediction is 0.594 angstroms.

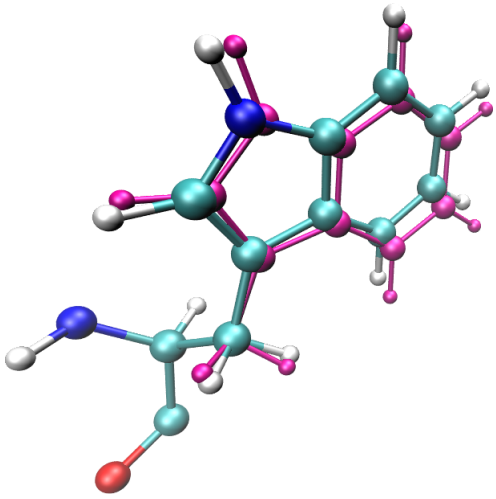
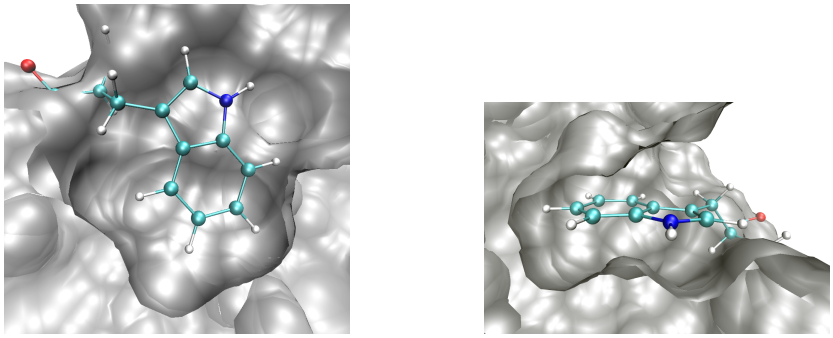


Figure 3.16: Native, colored by element, and predicted, magenta side chain conformation for tryptophan 43 of 1FCC. The root mean square distance of the predicted conformation to the native is 0.371 angstroms. The effect of the local protein structure on the conformation of this residue is examined in figure 3.17.



(a) (b)

Figure 3.17: The pocket of tryptophan 43 of 1FCC. Because of conformation of the neighboring protein structure this residue has very little conformational freedom, and any prediction which successfully locates the sidechain in the pocket will be reasonably close to the native state. The conformation predicted in these experiments was very similar, 0.371 angstroms, and is depicted superimposed with the native in figure 3.16.

3.4 Discussion

There are two necessary subproblems in accurately predicting the effect of mutations on local protein structure:

1. prediction of side chain conformations, and

2. accurately describing the energetics of the interactions.

In the work presented here we have shown that our current methods are able to accurately predict side chain conformations of mutated residues, discussed in 3.4.1. However, it seems that despite being able to differentiate between native and non-native side chain conformations, we are unable to correlate experimental changes in binding energy with computational predictions. Possible reasons for this discrepancy are addressed in 3.4.2.

3.4.1 Side Chain Prediction Accuracy

Our results indicate that we are able to successfully predict side chain conformations on protein interfaces in the majority of cases examined. Specifically, 29 of 32 side chains predicted over four chains in three independent structures are within 1.5 angstroms of the native conformation, with a significant number of side chains predicted closer to the native crystal conformation than the resolution of the crystal structure. This indicates that both the sampling performed here and the energy model are sufficiently extensive and accurate to reproduce the native conformation, which has been used as a standard metric of success in many previous studies, e.g. loop predictions [Jacobson *et al.*, 2004; Rapp and Friesner, 1999; Zhu *et al.*, 2006; Sellers *et al.*, 2008] and side chain predictions [Jacobson *et al.*, 2002b; Jacobson *et al.*, 2002a; Zhu *et al.*, 2007]. One of the reasons that RMSD is so popular as a performance metric is the difficulty of obtaining experimental data which can be directly compared to experimental predictions. Binding affinity studies, especially alanine scanning experiments represent a wealth of data that might be used in training more accurate next generation molecular mechanics energy functions.

3.4.2 Energetic Correlation with Experimental Data

We found that, despite predicting side chain conformations approximately correctly, there was generally no correlation between our computed $\Delta\Delta G$ and the experimental $\Delta\Delta G$ from the alanine scanning database, as shown in figures 3.2, 3.5, 3.9, and 3.12.

Experiments by other groups have demonstrated some success in correlating computational $\Delta\Delta G$ with experimental $\Delta\Delta G$ [Kortemme *et al.*, 2004]. Some of these experiments have made use of energy models which are largely similar to the one implemented in the PLOP program. Despite this, we did not find a significant correlation between experimental data and predictions with respect to $\Delta\Delta G$. It is possible that the interactions at a protein-protein interface are somehow different than the intramolecular interactions which have constituted the majority of the training sets used to develop the PLOP energy model.

3.4.3 Future Directions

Because the sampling introduced in these experiments is very modular in nature it would be possible to modify the sampling procedure used or even specify the sampling method to be used in the input file. Testing the performance of the energy model on protein-protein surfaces and classifying the errors would be a necessary step in improving the correlation between predicted and experimental $\Delta\Delta G$. It would also be very enlightening to be able to compare the performance of the PLOP sampling methods using a known energy model implemented in a different molecular mechanics toolkit.

Chapter 4

Prediction of P450 Sites of Metabolism

4.1 Introduction

The most common method of drug clearance among currently prescribed drugs is metabolism, which is the primary method of clearance for approximately 75% of the top 200 most commonly prescribed drugs in the United States [Williams *et al.*, 2004]. Cytochrome p450 is critical to drug metabolism, being active in approximately 75% of drugs which are cleared in this method [Guengerich, 2007]. Of the human isoforms of P450, Cytochrome P450 2D6 (CYP2D6) is frequently involved metabolism of xenobiotics [Williams *et al.*, 2004], there are also high resolution crystal structures available for CYP2D6 [Rowland *et al.*, 2006] and thus it was used as a test case for this study. As covered in 1.1.2.3, accurately predicting absorption, distribution, metabolism, and excretion, characteristics of drug compounds can be a critical determining factor in determining drug efficacy, performance in clinical development stages, and the overall costs of bringing new drugs to market. Because of the ubiquity of P450 in metabolic reactions of drugs, there is no other single enzyme family as significant to determining ADME as P450.

The general form of the reaction most frequently catalyzed by P450 is

$$\mathrm{RH} + \mathrm{O_2} + \mathrm{NADPH} + \mathrm{H^+} > \mathrm{ROH} + \mathrm{H_2O} + \mathrm{NADP^+} \tag{4.1}$$

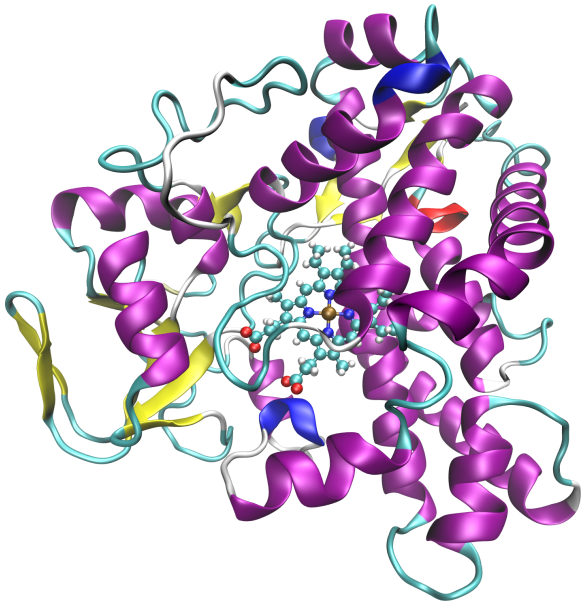


Figure 4.1: The structure of cytochrome P450, taken from PDBid 1JFB, shown in cartoon representation. The bonded heme group, shown as ball and stick model, is visible in the center. The brown iron atom is chelated by four deep blue nitrogen atoms.

The specific locations of sites of metabolism (SOM) on small molecules can have a profound effect on the ADME characteristics of a small molecule. Some cancer drugs such as epipodophyllotoxins, ifosphamide, tamoxifen, taxol and vinca alkaloids, are converted into their active states by oxygenation at specific locations by P450 [Kivisto *et al.*, 1995]. P450 is the body's primary defense against toxicity, usually catalyzing the conversion of toxic compounds into harmless products [Gonzalez, 2005; Guengerich, 2001]. However in certain cases, such as acetaminophen, it is possible for P450 to convert a harmless reactant into a toxic product [Chen *et al.*, 1998], although usually these compounds would be eliminated during the clinical trial stages. Additionally the different metabolites of a compound may be differentially cleared by the body having significant effects on bioavailability. Because of the costs associated with testing ADME parameters in live organisms accurate computational predictions can significantly decrease both costs and times associated with drug development.

Because of its central role in drug metabolism P450 has already been a subject of a number of studies attempting to predict sites of metabolism and chemical metabolites [Afzelius

et al., 2007]. A number of different classes of methods for predicting sites of metabolism by P450 have been developed. Broadly speaking these can be classified into: quantitative structure-activity relationship (QSAR) based, pharmacophore-based, structure-based (docking), reactivity-based, and rule-based methods [Cruciani *et al.*, 2005]. Rule based and pharmacophore based methods make predictions based on a subset of the drug structure, and it is possible for elements of the drug far from a possible site of metabolism to either prevent or promote metabolism at that location. QSAR based approaches work best when the set of reactions being catalyzed are very similar, however P450 catalyzes a very broad range of reactions so these approaches are likewise somewhat limited in the case of P450. Reactivity based methods are both very expensive to compute, being unsuited for screening a large database and do not take into account the structure of the P450 isoform [Singh *et al.*, 2003; Chen *et al.*, 1997; de Visser *et al.*, 2002]. MetaSite, an approach which makes use of structural information of both the ligand and the P450 isoform process has achieved a 84.3% prediction accuracy (296 of 351 total sites of metabolism correctly predicted), and the primary site of metabolism is identified in the top 3 ranked sites in over 90% of cases [Cruciani *et al.*, 2005]. However the sampling of P450 conformations done by MetaSite is quite limited, pre-computing a number of low energy conformations and then docking the substrate into each of those.

We have developed a similar approach which provides significantly more through sampling of the P450 substrate complex. The new method, IDSite, makes use of the structures of both the P450 and the substrate as well as evaluating the intrinsic reactivity of the possible site of metabolism.

4.2 Methods

Prediction of sites of metabolism is a three stage procedure:

1. Initially a number of different ligand conformations are generated, and these are docked into a rigid protein, with soft VDW terms using Glide [Halgren *et al.*, 2004; Friesner *et al.*, 2004].

2. The docked conformations are refined using a Monte Carlo Minimization (MMC)

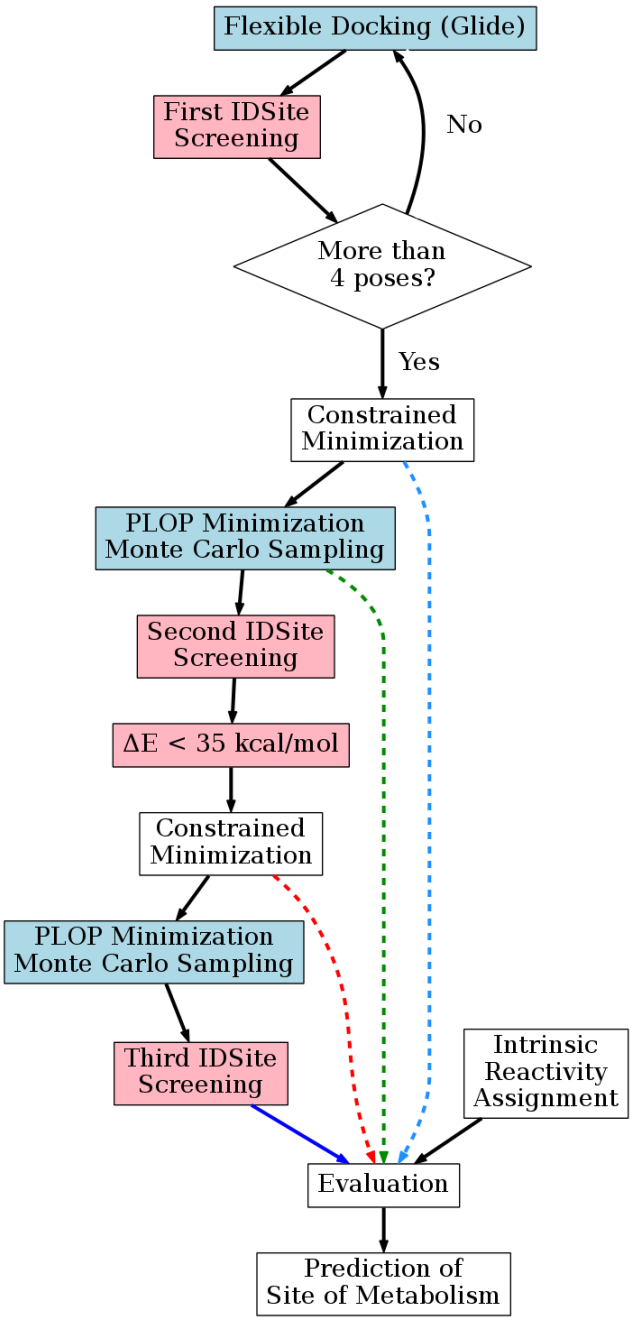


Figure 4.2: An overview of the entire IDSite procedure. The dotted lines represent abbreviated versions of the full procedure. Receiver operating characteristic graphs for the full version, and these abbreviated versions, are presented in 4.15. Series colors on ROC graphs correspond to arrow colors here.

approach which samples degrees of freedom in both the ligand and protein.

3. Refined conformations are classified into reactive site or non-reactive site on the basis of the energy of the refined conformations and the intrinsic reactivity of the site. [Li *et al.*, 2011b]

4.2.1 Docking

In the initial docking stage of the IDSite protocol Glide is used to generate a number of proposed docked conformations for each ligand. Glide (standard precision) is used to generate a number of different ligand conformations by sampling conformations of freely rotatable bonds and rings. A bounding box, which will be used for a grid search, is defined centered at the centroid of the ligand with an edge length of 10 angstroms. Because the crystal structure used for CYP2D6 (PDBID: 2F9Q) does not have a ligand, the centroid of residues Glu216, Asp301, Thr309, and Phe483 was used instead in this case. Because the steric clashes present in many proposed docked conformations can be relieved using a simple minimization procedure a reduced Van der Waals (VDW) radii are used in the docking stage for non-polar atoms. The VDW radii used for the P450 are scaled by a factor of 0.4, and the scaling for the ligand starts at 0.8. If an insufficient number of poses, in this case fewer than four, are found using these scaling factors for the radii the scaling of the ligand is stepped down until at least four poses are found. Additional filtering of possible high energy conformations was also skipped in order to ensure the greatest diversity of docked poses reached the refinement stage. The collection of docked poses are then clustered according to the RMSD of the ligand, and each pose is minimized. The top sixty ranked poses according to the Glide SP metric are retained screened using a number of different criteria. A hard sphere overlap criteria is used to remove poses with obvious steric clashes which were not removed during the minimization procedure. A conserved feature of CYP2D6 ligand complexes is a salt bridge with Glu216 or Asp301. In order to reduce sampling cost IDSite only considers structures with at least one hydrogen-bond donor within 4 angstroms of the centroid of these two residues and Ser304. The sphere defined by these residues is illustrated along with the bounding box used for sampling in Figure 4.3 A number of other rule based geometric screens are used to remove structures

which are unlikely to react. Structures meeting any of the following criteria:

1. The distance of the basic nitrogen to the ferryl oxygen is less than 5.0 angstroms;
2. The distance of the basic nitrogen to the negative charged oxygen (in Glu216 or Asp301) is greater than 5.5 angstroms;
3. More than 2 heavy atoms from the ligands are further than 16.0 angstroms away from the heme iron;
4. More than 1 heavy atom from the ligand are closer than 1.0 angstroms to the receptor;
5. More than 6 heavy atoms from the ligand are closer than 1.8 angstroms to the receptor;
6. No heavy atom in the ligand is within 5.0 angstroms to the heme iron;

are removed. If the number of structures at this point is too low, the VDW scaling factors of the non-polar atoms of the ligand are stepped down, and the process is repeated. If four or more poses are found at these point these poses are passed onto the next stage of the IDSite procedure, the Monte Carlo Minimization refinement stage.

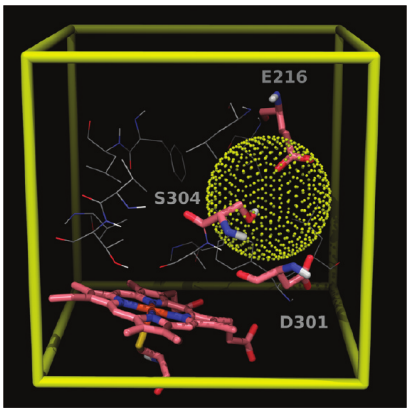


Figure 4.3: The bounding box used by Glide in order to generate the inital set of docked poses. The docking procedure also requires at least one hydrogen bond donor be found within 4 angstroms of the centroid of Glu216, Asp301, and Ser304 is also shown. The sphere representing this constraint is also shown.

4.2.2 Monte Carlo Minimization Refinement

Since the emphasis in IDSite sampling is efficient sampling of low energy conformations, as only the lowest energy conformations are passed on to the next stage of prediction, Monte Carlo Minimization, which provides more efficient sampling of low energy conformations, was used instead of a more traditional Monte Carlo simulation (see 1.2.2). The Monte Carlo Minimization sampling used by IDSite for refinement incorporates three different types of steps: side chain motions, rigid body transformations, and hybrid Monte Carlo simulations. For each Monte Carlo step one of three types of motions is selected according to the weighted probabilities, which are different for the two different PLOP sampling stages, see Table 4.1. Using the chosen method a new conformation is proposed and minimized

	PLOP Sampling Stage	
	First	Second
Number of Residues Sampled	12	40
Number of Structures Advanced to Next Stage	max(n*8,24)	max(n*20,60)
P(side chain step)	0.5	0.7
P(rigid body step)	0.1	0.2
P(HMC)	0.4	0.2

Table 4.1: The number of residues sampled as well as the number of structures advanced to the next stage from each of the sampling stages. Also, the relative probabilities of selecting each of the different sampling steps during a Monte Carlo minimization sampling stage.

before the Metropolis acceptance criteria (equation 1.1) is applied to the proposed state, using a temperature of 300 K. All atoms of all residues with any atom within 5 Angstroms of the ligand in the starting crystal structure were allowed to move during Monte Carlo moves, including the ligand itself.

During the minimization Monte Carlo sampling stages of the IDSite procedure artificial constraints are used to guide the sampling towards a transition state like conformation. These constraints create artificial bond or angle potentials which affect the minimization, but are not used in the Monte Carlo acceptance test. For each of the minimization Monte

Carlo sampling stages of the IDSite procedure two different sets of constraints are applied depending on the hybridization of the carbon atom at the possible site of metabolism, for a total of four possible different sets of constraints. In the first minimization Monte Carlo stage two constraints are applied:

1. The sulfur-iron-carbon angle is constrained to 145 degrees, with 20 degrees of "slack", or a flat bottom to the potential well (denoted as 145±20 degrees). The spring constant of this constraint is about 25 kcal/mol/degree^2, or ~40% the strength of a carbon-carbon-carbon angle.

2. A "dummy" oxygen atom is placed above the plane of the heme group, in the same position that it would occupy if an oxygen molecule was bound to the heme. This dummy atom has no interactions with other atoms, but is used as the anchor of a distance constraint for the carbon at the site of metabolism. The carbon-dummy oxygen distance is constrained to 2.5±0.5 angstroms. The spring constant of this constraint is 100 kcal/mol/angstrom^2, approximately 1/3rd the strength of a carbon-carbon bond.

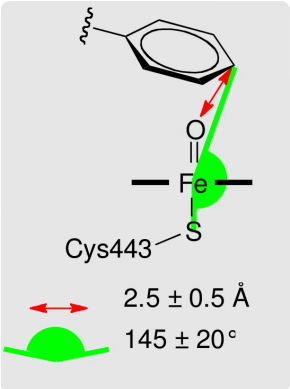


Figure 4.4: The constraints applied to sp^2 atoms during the constrained minimization and first minimization Monte Carlo sampling stage. The spring constant of the bond constraint (red arrow) is 100 kcal/mol/angstrom^2, and that of the angle constraint is 25 kcal/mol/degree^2. The oxygen atom depicted in this figure is a "dummy" atom and does not interact with any other atoms in the structure except through the constraint.

In the second minimization Monte Carlo sampling stage the constraints are different for sp^2 and sp^3 carbons. For sp^3 sites:

1. the hydrogen bound to the carbon at the possible site of metabolism is constrained to a distance of 1.25±0.1 angstroms and a spring constant of 20 kcal/mol/angstrom^2,
2. the carbon in question is constrained to 2.2±0.8 angstroms and a spring constant of 10 kcal/mol/angstrom^2,
3. the heme iron-hydrogen-carbon angle is constrained to 138±5 degrees and a spring constant of 20 kcal/mol/degree^2.

For sp^2 sites:

1. the carbon at the possible site of metabolism is constrained to 1.8±0.1 angstroms and a spring constant of 20 kcal/mol/angstrom^2,
2. both adjacent carbons are also constrained to the dummy oxygen atom, at a distance of 2.5±0.1 angstroms and a spring constant of 20 kcal/mol/angstrom^2, and
3. finally the hydrogen bonded to the carbon at the possible site of metabolism is constrained to the oxygen atom at a distance of 2.0±0.1 angstroms and a 20 kcal/mol/angstrom^2 spring constant.

As CYP2D6 forms a conserved salt bridge with the substrate with either glutamate 216 and aspartate 301 [Paine *et al.*, 2003], this was also incorporated as a constraint during the sampling stages. In the first sampling stage this salt bridge is enforced by introducing a harmonic constraint of 3.0±0.3 angstroms, between the basic nitrogen of the substrate and each of the side chain oxygen atoms in GLU216, ASP301 and SER304. The spring constants of this constraints are 15.0, 8.0 and 4.0 kcal/mol/angstrom^2 for GLU216, ASP301 and SER304 respectively. Additionally, an angle constraint is applied to each of the N-H-O angles, this is set to 150.0±30.0 degrees and has a spring constant of 5.0 kcal/mol/degree^2. In the second sampling stage four separate trajectories are calculated for each of the four carboxalate oxygens of GLU216, ASP301. In each trajectory a constraint of 1.9±0.1 angstroms is applied between the hydrogen attached to the basic substrate nitrogen and one of the

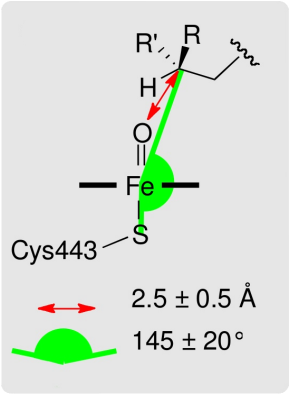


Figure 4.5: The constraints applied to sp^3 atoms during the constrained minimization and first minimization Monte Carlo sampling stage. The spring constant of the bond constraint (red arrow) is 100 kcal/mol/angstrom^2, and that of the angle constraint is 25 kcal/mol/degree^2. The oxygen atom depicted in this figure is a "dummy" atom and does not interact with any other atoms in the structure except through the constraint.

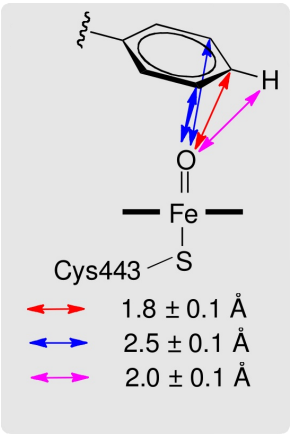


Figure 4.6: The constraints applied to sp^2 atoms during the constrained minimization and second minimization Monte Carlo sampling stage.

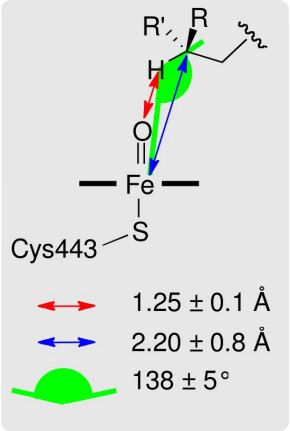


Figure 4.7: The constraints applied to sp^3 atoms during the constrained minimization and second minimization Monte Carlo sampling stage.

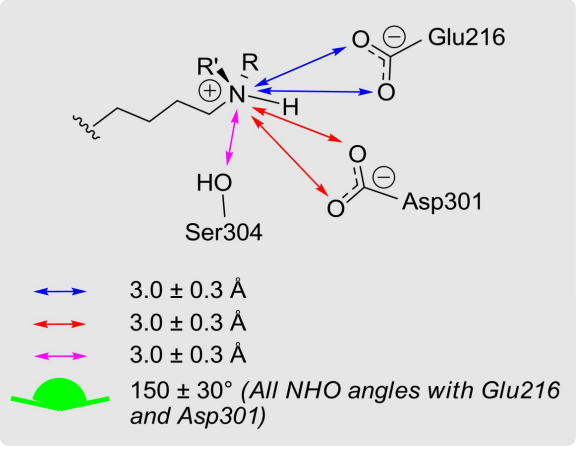


Figure 4.8: The constraints applied to the salt bridge region of CYP2D6 during the *first* minimization Monte Carlo sampling stage.

four carboxalate oxygens. Additionally, the angle of the hydrogen bond is constrained to 168±12 degrees, with a spring constant of 5.0 kcal/mol/degree^2.

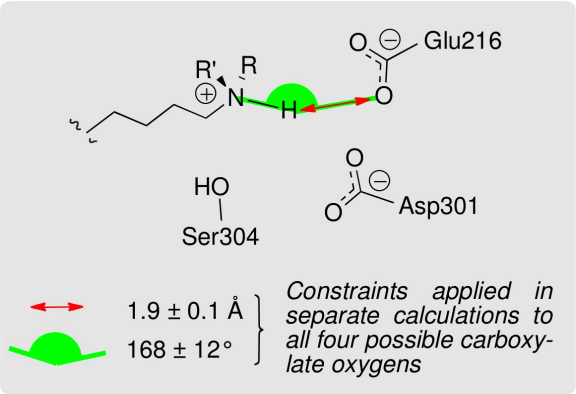


Figure 4.9: The constraints applied to the salt bridge region of CYP2D6 during the *second* minimization Monte Carlo sampling stage.

Three different methods of sampling, or moves, were implemented in order to sample different protein-ligand conformations.

1. Side chain motions seek to sample different conformations of the neighboring side chains and flexible moieties attached to a static constrained central core of the ligand. Because side chain conformations are locally highly correlated, neighboring side chains are sampled simultaneously. First a number of side chains to be sampled is selected, for this study the possible cluster sizes were from one to three residues, all with equal probability. Once a cluster size is selected a single side chain is selected, at random. Until the desired cluster size is reached additional residues are selected, and rejected if the β carbon is greater than 6 angstroms from any of the β carbons of the residues already included in the cluster.

   Once the side chains have been selected three different possible moves are possible for each side chain, these are:

   (a) rotamer library based moves

   (b) random perturbations of a single dihedral

   (c) random perturbation of all dihedrals

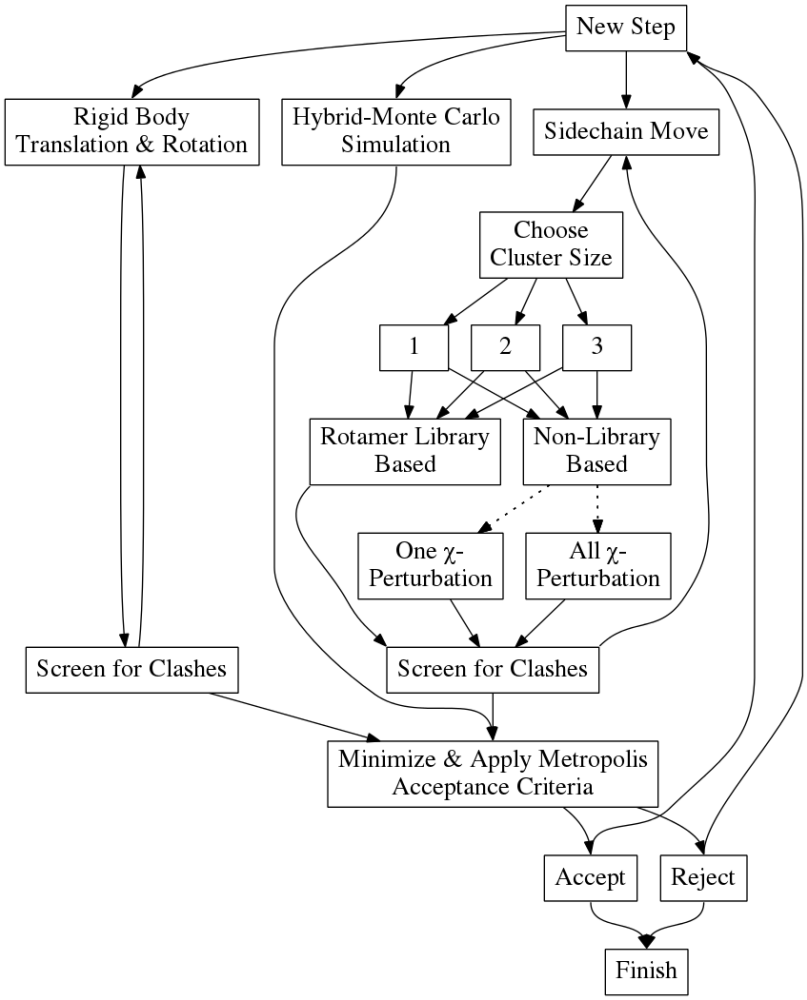


Figure 4.10: An outline of the Monte Carlo minimization refinement stages in PLOP.

For 75% of side chains, the rotamers are replaced with rotamers drawn at random from a high resolution rotamer library. The rotamer states for protein residues is the same library used in other place in PLOP, constructed from a set high resolution protein X-ray crystal structures [Xiang and Honig, 2001]. The rotamers for the ligand are constructed by sampling rotable bonds at 10 degree resolution and screening this set for steric clashes. In additon to the rotamer selected a small amount of "noise" which is random, and less than the resolution of the side chain library, is added to each rotamer.

For the remaining 25% of side chains, moves are distributed equally between perturbations of a single dihedral and perturbations of all dihedrals. The perturbation for each dihedral is constructed of both a small and a large perturbation. The small perturbation is chosen at random in the range [-d,d] where d is the resolution of the sidechain library, in this case 10 degrees. This term corresponds to the "noise" term in the case of rotamer moves. The large perturbation is a multiple of 60 degrees, which corresponds to the minima of both sp2 and sp3 carbons, though the null large perturbation is weighted more heavily. The effect of the union of these two terms is that dihedrals which are expected to yield higher acceptance probabilities are selected more frequently, though with the large perturbations this is balanced with large changes in conformation.

After a set of dihedral angles has been proposed for all residues in the cluster the atomic locations are updated. Finally, before testing for acceptance of the new coodinates a screen is a applied. A reduced radius, 0.6x scaling factor, hard sphere overlap screening criteria is applied for all possible side chain moves before a possibly costly minimization step. If the proposed structure does not pass the screening stage another side chain step is performed to obtain a new proposed conformation.

2. Rigid body translation and rotation were also implemented for noncovalently linked moieties, such as ligands. Rigid body steps are implemented as a multiple time scale Monte Carlo simulation. In inner steps a more lenient screening criteria is applied, and only short range interactions are updated. An inner step consists of a translation and

a rotation. For translations the ligand is translated a random distance between 0 and 0.5 angstroms in a direction is selected at random. Additionally, the ligand is allowed to rotate, this is accomplished by choosing another random vector through the center of mass, and rotating by an angle distributed at random between -60 and 60 degrees. Coupling rotations and translations allows sampling of concerted movement of the ligand. During inner steps the atomic radii are scaled by 0.7 when testing for atomic collisions. 1000 attempted inner steps are performed for each outer step. In the outer Monte Carlo step the full energy is calculated and the unscaled Lennard-Jones radii are used.

This multi-scale sampling method increases the ability to escape from tight spacial bottlenecks. This increases the conformational freedom and therefore sampling of the inner steps, at a cost of decreasing the acceptance probability in the outer loop. However, because a minimization was performed before testing for acceptance the acceptance criteria for rigid body steps was still 0.1 to 0.4 depending on the size of the ligand. Each rigid MC step consisted of 1000 inner steps, and only one outer step, meaning that only one minimization occurred each time rigid body Monte Carlo was selected as the move step. The rigid body move usually takes 20 to 40 seconds per move.

3. The Hybrid Monte Carlo (HMC) [Duane *et al.*, 1987] step is a velocity verlet molecular dynamics simulation. This simulation allows all atoms in both the ligand and residues containing atoms withing 5 angstroms of the ligand to move. Initial velocities are taken from a Maxwell-Boltzmann distribution at 900 K. Bonded and short range interactions evaluated every 1 nanosecond inner time step, and long range potentials are assumed to be fixed over inner steps. Five inner steps compose each outer HMC step. In the outer step the molecular surface, long range interactions and, Born alphas are updated before computing the energy and applying the Metropolis acceptance criteria at a temperature of 900 K after each MD run. Taking up to 15 minutes per move, the HMC is the most expensive among all three types of moves in PLOP.

4.2.3 Evaluation

Both a parameterized and an unparameterized model were used to classify potential sites of metabolism. IDSite makes the assumptions that all intermediates before the rate determining step are at equilibrium [Wang *et al.*, 2007], that hydrogen abstraction is the rate limiting step for hydroxilation of aliphatic carbons and electrophilic attack is the rate limiting step for hydroxilation of aromatic rings [Guengerich, 2001; Shaik *et al.*, 2005]. With these assumptions the rate of metabolism at each possible site of reaction is affected by the free energy of binding in order to put that site in the site of reaction, as well as the free energy barrier of rate determining step, or

$$\Delta G_{\text{total}} = \Delta G_{\text{binding}} + \Delta G_{\text{barrier}} \tag{4.2}$$

The $\Delta G_{\text{binding}}$ above is calculated using a PLOP evaluation of the refined pose. The intrinsic reactivity for the system is computed from DFT calculations on a simplified system, replacing the heme with a methody radical, and using a linear relationship between IR(heme) and IR(methoxy radical) to estimate the true reactivity for the heme system.

$$\text{IR(heme)} = 1.117 * \text{IR(methoxy radical)} + C \tag{4.3}$$

Since this constant C is identical for each state it has no effect on the relative differences in ΔG_{site} or the relative rate of metabolism at possible sites.

$$E = \langle 1.117 * \text{IR(methoxyradical)} + C + E_{\text{TS}} \rangle - kT\ ln(N_H) \tag{4.4}$$

Since the ligand is forced to assume a different conformation in order to react, the energy of this transition state conformation, E_{TS}, is also computed using PLOP. As the relative abundance of different metabolites is determined by differences in ΔG per site rather than absolute reactivities, the constant in equation 4.4 does not affect which metabolites are produced. A site of possible metabolism is classified as positive if it is observed in greater than 0.1% yield, which corresponds to a $\Delta\Delta G$ of ~4.75 kcal/mol between the most favored state and the cutoff for negative predictions.

The second classifier is similar however:

1. a different constant is used to estimate IR(heme) from IR(methoxy radical), namely 1.071,

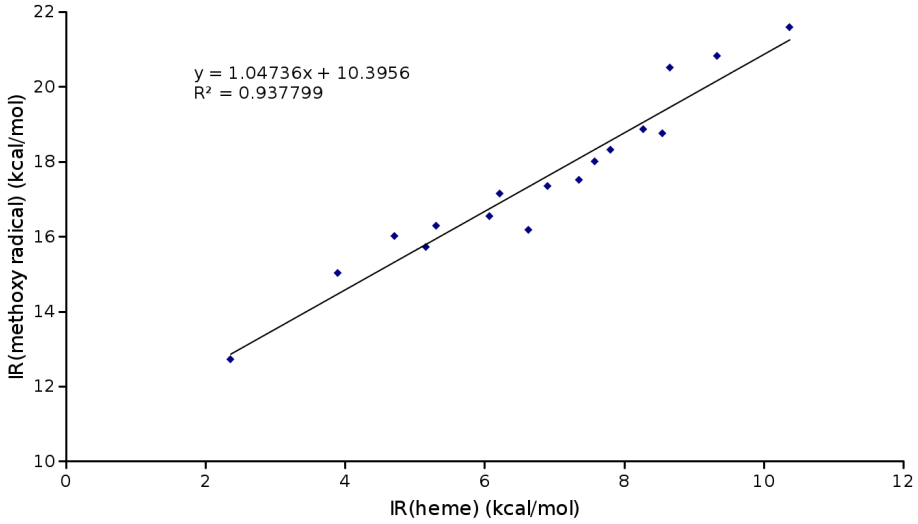


Figure 4.11: The linear relationship between the calculated intrinsic reactivity of the methoxy radical complex and that of the heme complex. Adapted from [Li *et al.*, 2011b] with minor correction. In the original manuscript the slope of the regression was reported as 1.117 and that number was used throughout. This difference should not significantly affect the physical IDSite classifier results, and does not affect the results of the fit model. In the rest of this text the value from the original publication of 1.117 will be used.

2. if the binding energy of the transition state complex of a pose is within 5.26 kcal/mol of the lowest pose, it is set to the binding energy of the lowest pose. Otherwise the difference is scaled by 0.58,

3. and the cutoff for an active prediction is changed from 4.75 kcal/mol to 1.46 kcal/mol.

These parameters were decided upon by maximizing $\frac{\text{true positives}}{(\text{false positives}+\text{false negatives})}$ on a training set of 36 compounds.

Model compound	Site of Metabolism	Heme model (kcal/mol)	Methoxy model (kcal/mol)
Benzene		20.51	8.66
Anisole	Ortho-	16.29	5.31
	Meta-	18.76	8.55
	Para-	16.01	4.71
	Beta-	16.18	6.63
Dimethylether		15.03	3.9
Dimethylanisole	Meta-	16.54	6.07
	Para-	17.51	7.35
Ethane		21.58	10.37
Ethanol	1	12.73	2.36
	2	17.35	6.9
Propane		18.31	7.8
Toluene	Ortho-	17.15	6.22
	Meta-	18.86	8.27
	Para-	18	7.58
	Alpha-	15.72	5.16
t-Butylebenzene	Beta-	20.82	9.33

Table 4.2: DFT calculated values for internal reactivity of various compounds with either methoxy radical (compound I) or heme system. Correlation between these values is illustrated in Figure 4.11.

4.3 Results

Both physical and fitted IDSite were able to acheive promising results predicting CYP2D6 sites of metabolism.

The physical model correctly identified 68 of 82 active sites of metabolism for a sensitivity of 0.829. For inactive sites this model correctly identified 1054 of 1075 inactive sites with a specificity of 0.980. The fit model performed similarly, and even slighly better identifying 52 of 57 sites of metabolism (sensitivity of 0.912) in the training set and 25 of 25 in the test set (sensitivity of 1.0).

$$\text{sensitivity} = \frac{TN}{TN + FP} = \frac{\#\text{ of true sites of non-metabolism identified}}{\#\text{ experimental sites of non-metabolism}} \tag{4.5}$$

$$\text{sensitivity} = \frac{TP}{TP + FN} = \frac{\#\text{ of true sites of metabolism identified}}{\#\text{ experimental sites of metabolism}} \tag{4.6}$$

The fit model also correctly identified 709 of 717 inactive sites in the training set (specificity of 0.989) and 352 of 358 inactive sites in the test set (specificity of 0.983). As the performance of the fit model is similar to that of the physical model, it does not appear that the fit model is overparameterized to the training set. Specific results for both models are presented in Tables 4.3 and 4.4, and the specific sites identified by both models as well as experiments are illustrated in Figures 4.13 and 4.14.

We believe that the parameters help account for some degree of noise in the molecular mechanics calculations. The scaling of the binding energy difference, either to zero inside a window about the minimum energy pose, or by a factor of 0.58 decreases the relative weight of the molecular mechanics contribution relative the quantum contribution to the classifier. This might imply that some sites are not being classified as active because they are not in the lowest energy conformation around the docked pose, suggesting that additional molecular mechanics sampling might further improve results. However as will be discussed later, the molecular mechanics stage already dominates the total time necessary for an IDSite prediction, and the current molecular mechanics procedure takes about 450 hours.

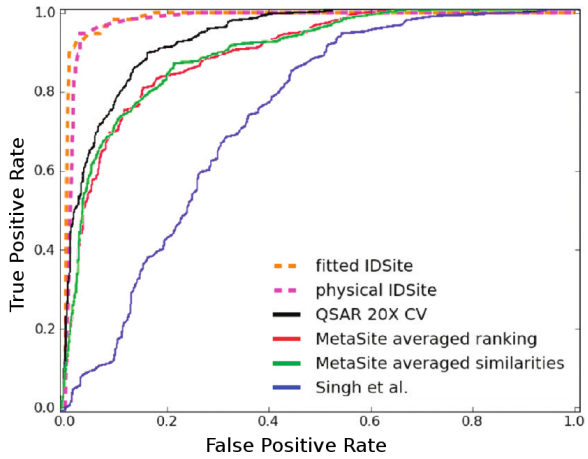


Figure 4.12: A comparison of the performance of IDSite with a variety of other methods of predicting P450 sites of metabolism. IDSite obtains the best performance, followed by a quantitative structure-activity relationship based method [Sheridan *et al.*, 2007]. Adapted from [Sheridan *et al.*, 2007].

		Physical			Fitted		
Compound #	Compound	TP	FP	FN	TP	FP	FN
1	4-methoxyamphetamine	1	0	0	1	0	0
2	Amitriptyline	2	2	0	2	0	0
3	Aprindine	4	0	1	5	0	0
4	Brofaromine	1	0	0	1	0	0
5	Bufuralol	0	1	1	1	0	0
6	Carvedilol	1	0	2	2	0	1
7	Cinnarizine	0	2	1	0	2	1
8	Clomipramine	1	0	1	1	0	1
9	Codeine	1	0	0	1	0	0
10	Desipramine	2	0	0	2	0	0
11	Dextromethorphan	1	0	0	1	0	0
12	Dihydrocodeine	1	1	0	1	0	0
13	Ethylmorphine	1	0	0	1	0	0
14	Flunarizine	1	0	0	1	0	0
15	Fluperlapine	1	0	0	1	0	0
16	Hydrocodone	1	0	0	1	0	0
17	Imipramine	2	0	0	2	0	0
18	Indoramine	1	0	0	1	0	0
19	MDMA	1	0	0	1	0	0
20	Methamphetamine	1	0	0	1	2	0
21	Methoxyphenamine	2	0	0	2	0	0
22	Metoprolol	1	0	1	2	0	0
23	Mexiletine	2	0	1	2	0	1
24	Mianserin	1	0	0	1	0	0
25	Mirtazapine	0	1	1	1	1	0
26	Nortriptyline	1	1	0	1	0	0
27	Ondansetron	2	0	0	1	0	1
28	Paroxetine	1	0	0	1	0	0
29	Perhexiline	2	0	0	2	0	0
30	Propafenone	1	1	0	1	1	0
31	Propranolol	2	2	0	2	1	0
32	Tamoxifen	1	0	0	1	0	0
33	Terfenadine	3	0	0	3	0	0
34	Tiracizine	1	2	0	1	1	0
35	Tropisetron	2	0	1	3	0	0
36	Venlafaxine	1	0	0	1	0	0
	Total	47	13	10	52	8	5

Table 4.3: Results of physical and fitted IDSite on training set of 36 compounds.

		Physical			Fitted		
Compound #	Compound	TP	FP	FN	TP	FP	FN
37	Atomoxetine	0	1	1	1	2	0
38	Bicifadine	1	2	0	1	0	0
39	Bupranolol	1	0	0	1	0	0
40	Carteolol	1	1	0	1	0	0
41	Chlorpromazine	1	0	0	1	0	0
42	EMAMC	1	0	0	1	0	0
43	Encainide	1	1	0	1	1	0
44	Harmaline	1	0	0	1	0	0
45	Harmine	1	1	0	1	1	0
46	Ibogaine	1	0	0	1	0	0
47	MAMC	1	0	0	1	0	0
48	MMAMC	1	0	0	1	0	0
49	MOPPP	1	0	0	1	0	0
50	Oxycodone	1	0	0	1	0	0
51	Spirosulfonamide	2	0	0	2	0	0
52	Timolol	2	0	2	4	0	0
53	Tolterodine	0	1	1	1	1	0
54	Tramadol	1	1	0	1	1	0
55	Tyramine	2	0	0	2	0	0
56	Zotepine	1	0	0	1	0	0
	Total	21	8	4	25	6	0

Table 4.4: Results of physical and fitted IDSite on a test set of 20 compounds. Note that for the physical model there is no training performed so results in the text are presented in a unified fashion for the training and test set.

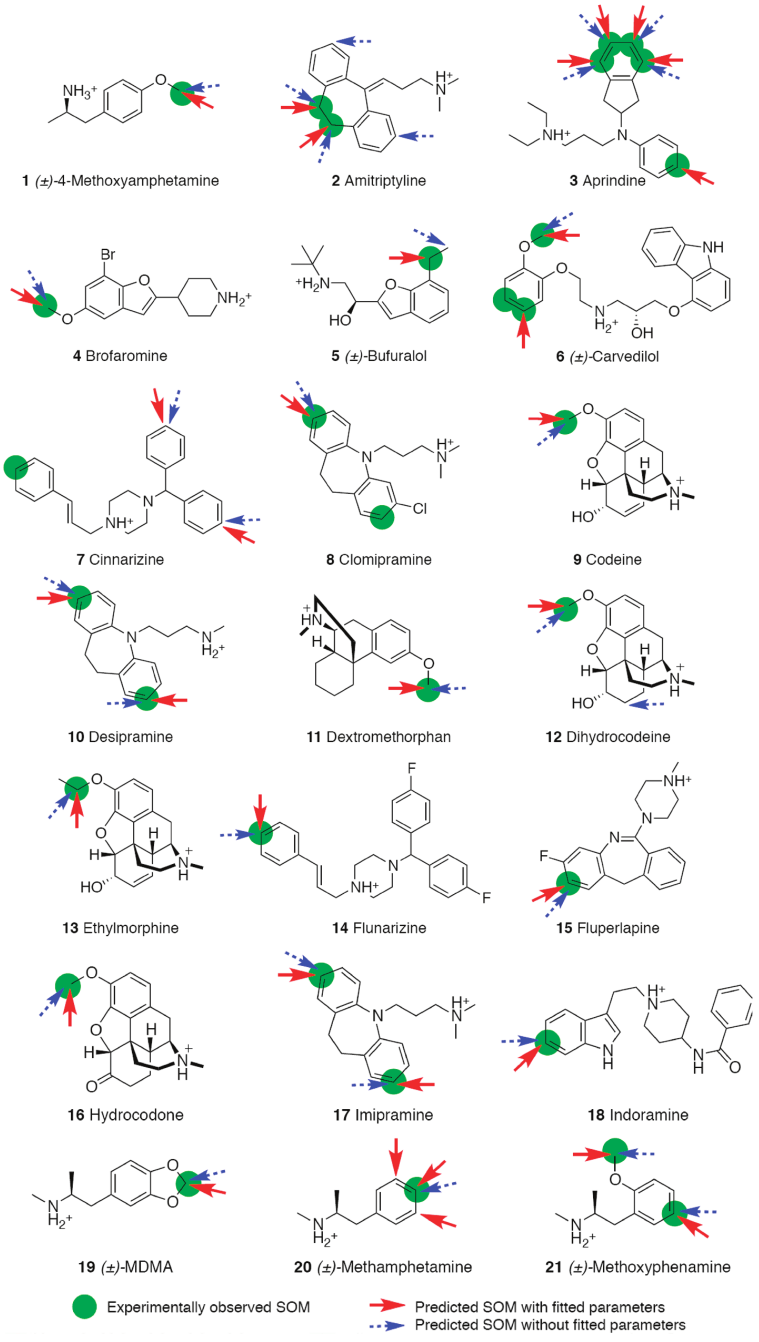


Figure 4.13: Physical and fitted IDSite predictions of sites of metabolism on the training set.

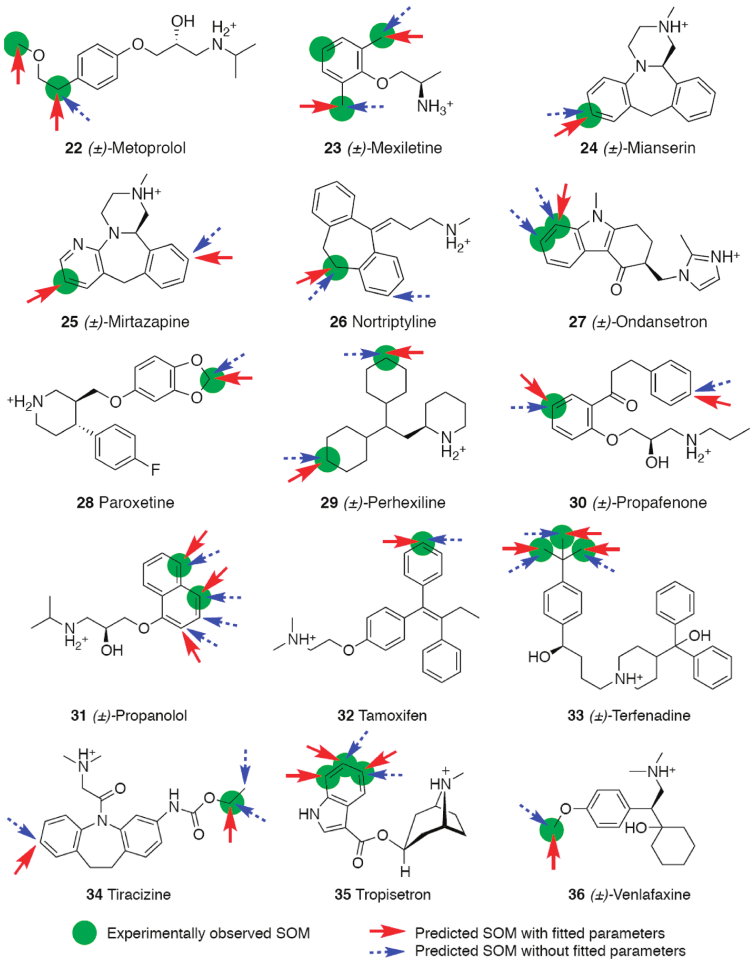


Figure 4.13: (continued)

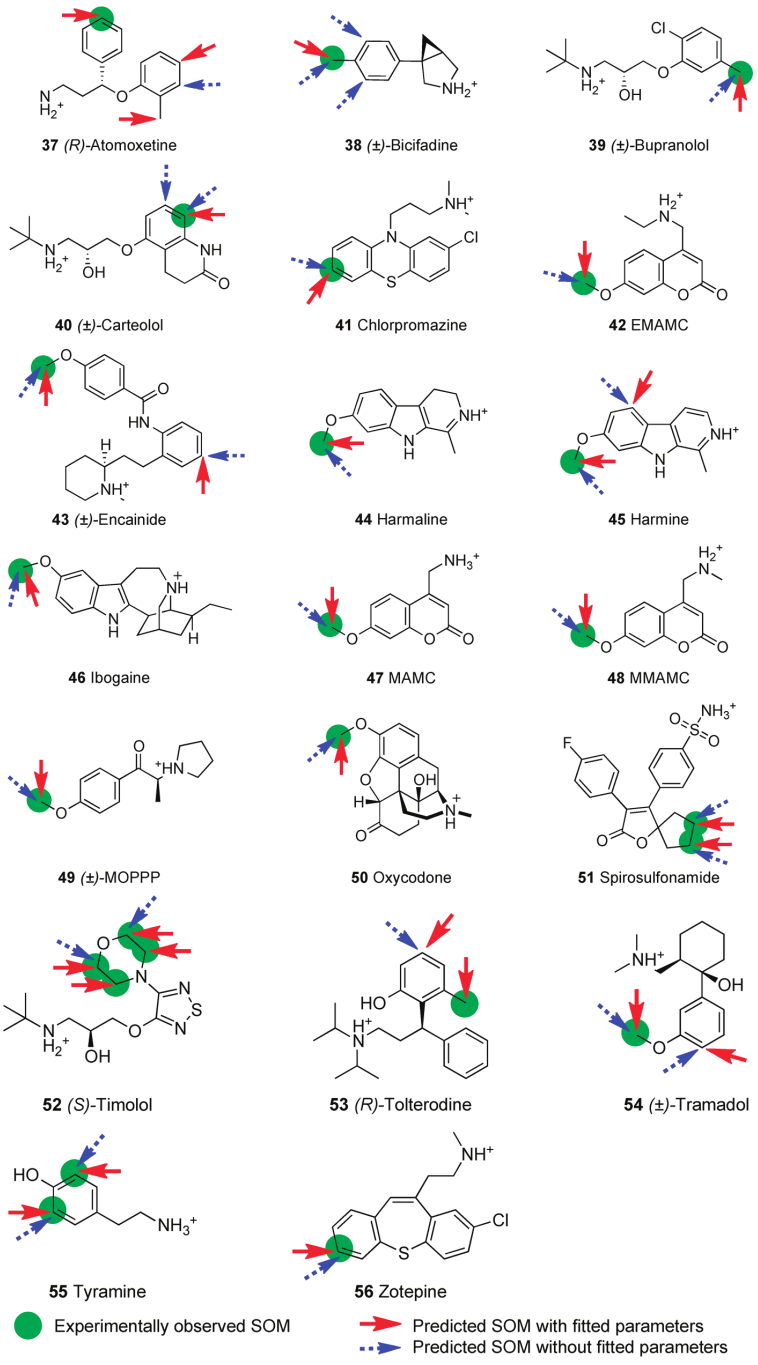


Figure 4.14: Physical and fitted IDSite predictions of sites of metabolism on the test set.

4.4 Discussion

The physical score performs nearly as well as the fitted score indicating that the model is not over-fit to the training set. Figure 4.12 presents a comparison of both IDSite models (physical and fit) to a number of other methods for prediction of P450 sites of metabolism. Though the data sets used are not identical they are similar and overlapping for some compounds. At all sensitivities IDSite is clearly the best performing protocol, with the physical and fit model performing very nearly identically.

At 90% sensitivity, the QSAR method included roughly 20% false positives, and MetaSite included 40% false positives. At the same sensitivity IDSite has a ~1% false positives rate. This represents a significant improvement in prediction accuracy that has the potential to have a large impact on drug development.

4.4.1 Significance of Sampling Stages

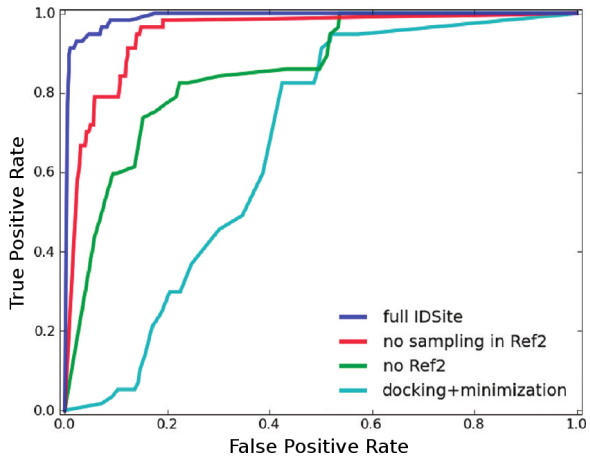


Figure 4.15: The effect of additional sampling on prediction of site of metabolism by P450. The light blue series describes only performing the initial Glide docking stage followed by minimization. The green series is obtained by using the set of structures obtained in the first minimization Monte Carlo sampling stage. The red series is obtained by screening the structures obtained in the first sampling stage, and minimizing these structures using the constraints specified in Figures 4.6 and 4.7. The blue series makes use of the entire IDSite procedure. The color scheme of these series corresponds to the colors of edges in Figure 4.2.

As shown in Figure 4.15 at a given specificity, sensitivity increases with additional sampling. For some compounds there are either no conformations predicted via the docking procedure where the site of metabolism is sufficiently close to the heme oxygen to react or the scoring metric used by the docking procedure prefers conformations which are far from the reactive site over more reactive conformations. The sampling stages, and constraints are useful in guiding these initial docked poses into more active conformations.

The trajectory followed during the minimization Monte Carlo sampling stages differs appreciably for small and large ligands. For smaller ligands, in many cases the poses after the first constrained minimization were among the lowest energy poses explored. For about 40% of these small ligands (24/56) the same predictions were obtained after the first minimization Monte Carlo sampling stage as after the full IDSite procedure. However, for large ligands, the constraints and sampling continued to decrease the energy over the course of the sampling stages, indicating that especially for large compounds the second refinement stage is critical to obtaining accurate predictions.

4.4.2 Induced Fit Effects

The catalytic site of cytochrome P450 is flexible, changing conformation to accommodate a large variety of different substrates [Li *et al.*, 2004; Scott *et al.*, 2004]. The Monte Carlo sampling stages reflect this flexibility allowing the substrate to settle into the active site, guided by harmonic constraints, and allows the side chains to alter their conformations to better accommodate and interact with the ligand. Side chain dihedrals were found to change less for smaller ligands, illustrating that less rearrangement of the active site was necessary in order bind the substrate in an active conformation. In larger ligands, some side chain dihedrals changed by larger amounts, especially for bulkier amino acids such as phenylalanine, with Phe120 and Phe483 changing by up to 40 and 60 degrees respectively for some ligands.

4.4.3 Balancing Structural Contribution and Reactivity

As reflected in equation 4.2, the binding energy and intrinsic reactivity are the determining factors in predicting sites of metabolism. The IDSite procedure evaluates and uses both of

these quantities in classifying possible sites of metabolism. The combination of structural information and information about the intrinsic reactivity of a site allows IDSite to correctly predict cases which would not be possible using either of these methods alone.

One case of such a prediction is nortriptyline, since the two sites on the seven-membered aliphatic ring are difficult to distinguish only with their intrinsic reactivity, as they are almost equally reactive. However, experiments show that only the (E)-10 site of nortriptyline is metabolized [Linnet and Olesen, 1997]. In IDSite, the poses with the (Z)-10 site close to the ferryl oxygen are all at least 10 kcal/mol higher in energy compared to the poses with the (E)-10 atom close to the ferryl oxygen. Such an energy gap is large enough for IDSite to correctly determine the (E)-isomer as the only metabolite. While structural effects are therefore clearly very important to determine nortriptylene's site of metabolism, the intrinsic reactivities also play a key role. This is again nicely illustrated with the example of nortriptyline, where a simply structure based method (without considering intrinsic reactivities) would predict the site of metabolism as being an aromatic hydroxylation due to the favorable energy of the corresponding poses. Therefore, IDSite is able to correctly balance the subtle effects stemming from intrinsic reactivity and structural fit.

Another case which illustrates the importance of balancing reactivity and structural information is brofaromine. Experiments show that the major site of metabolism by CYP2D6 is O-demethylation [Feifel *et al.*, 1993]. The intrinsic reactivity of the site of metabolism (4.7 kcal/mol) is very close to those of sites on the aromatic rings (non-sites of metabolism, 3.3 to 4.9 kcal/mol). Due to the receptor geometry, it is impossible for the atoms on the furan ring to get close to the ferryl oxygen while still attaining the salt bridge with either Glu216 or Asp301. Therefore, no poses which which would have predicted reaction on the furan ring satisfied the geometric constraints and those sites were correctly predicted as sites of non-metabolism. Although poses which satisfied the geometric constraints and placed the benzene ring in the metabolic site, these poses are strongly disfavored energetically by more than 20 kcal/mol. This indicates that taking the interactions between the ligand and the receptor into account, IDSite is able to make the prediction of the site of metabolism for brofaromine in good agreement with the experimental observation.

Chapter 5

Other Improvements in the Protein Local Optimization Program

5.1 Regression Testing

A number of common problems inherent in large programming projects are identifying bugs or regressions. A common means of minimizing the cost of maintaining a project is regression testing. Regression testing identifies a number of test cases which ideally provide good coverage of the code base and are able to identify changes which may have adversely affected established features of the program [Wong *et al.*, 1997]. When testing is performed continuously the time spent testing can account for as much as half of all development time [Leung and White, 1989]. However, in cases where this testing is neglected large software projects can easily reach the point where it is no longer feasible to repair or reverse-engineer [Weide *et al.*, 1995].

We have implemented a regression testing framework for the Protein Local Optimization Program with the goal of making the program more maintainable. Using this tool, after each modification a number of tests are performed and the results are compared with results from a known reference. All of this is performed in parallel, and without the users interaction. This allows good code coverage without requiring an unwieldy amount of time.

Tests can be submitted as new features are developed, hopefully maintaining good code coverage. Each test file contains three pieces of information:

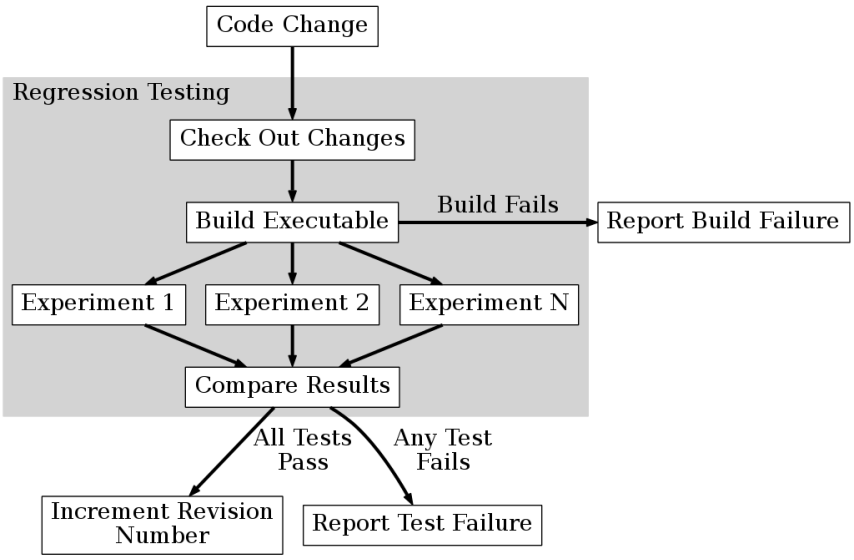


Figure 5.1: An overview of the regression testing procedure.

1. an experiment to be run,
2. a set of regular expressions matching the values to be compared between code changes,
3. for numeric comparisons, a threshold, or tolerance for change in a test before it is classified as a failure.

These test files are submitted to the version control system (CVS, git). The regression tester checks out a copy of these tests along with the code necessary to build an executable. The executable is then built, logging any warnings or errors, and assuming the build is successful the experiments described in the test files are run saving all output files in an archive. The results for each test, identified using a set of patterns in the test file, are then compared to the last "passing" set of results in the archive (if there is no old copy of a result, as in the case of testing a new feature, that test is assumed to pass). PLOP uses the major.minor.revision numbering scheme, and after a set of tests runs to completion and successfully reproduces the previous results, within the threshold, the revision number is changed. If a qualitative change in behavior is desired, i.e. larger than the threshold defined by the test files, incrementing the version number will allow the test to pass. This guarantees that results between two different revisions differ by no more than the threshold

for each of the tests specified, which theoretically allows one to reproduce results simply by using the same major and minor version.

At the present functionality tested includes: loop predictions, side chain prediction, truncated Newton minimization, a number of different energy functions, protonation state prediction, the minimization Monte Carlo sampling procedure described in 4.2.2, loading from sequence, retrieval of structures from the PDB, the knowledge based dihedral potential described in 5.3, and every space group present in the PDB as of 2012. An approximate performance based comparison is also included, which ensures that the time required for each test increases by no more than 10% between successive revisions.

Presently 159 tests are performed which takes about three hours. However, because the tests can be performed in parallel with a sufficient number of CPUs available the time is limited by the execution of the longest experiment. Using an eight core machine the total execution time for these tests is about 20 minutes.

Although in a perfect world all changes would be tested throughly enough such that they no adverse affects on disparate parts of the program, this is not possible due to practical considerations. This allows developers to spend their time working on relevant and related parts of code, and leave tedious and slow testing to an automated process.

This was originally implemented because fragmentation had led to an inability to reproduce results between different versions of the program. By performing a scan over time the specific changes which caused the difference between two versions were able to be quickly identified and addressed. Performing the same task by hand might have required reviewing all the code changes over a three year period.

¿¿¿¿¿¿¿ bf266b90d7fa3f6c1d52947d47038b58b3503865

5.2 Small Molecule Library

PLOP was originally developed to predict protein side chain and loop conformations. Many of the studies performed using this program have explicitly avoided cases in which proteins were near metal ions or lignds. Generally cases in which a charged ligand is closer than 6.5 angstroms, or a neutral ligand is closer than 4 angstroms is rejected [Goldfeld *et al.*, 2011;

Miller *et al.*, 2013]. This is partially because energy models have been trained using proteins, but also because for many ligands parameters for the various terms of the energy function are not available. At present the PDB contains 15,612 small molecule ligands, and previous experiments have used a limited library of 4,321 small molecules. Additional ligands have either been

1. removed from structures, under the assumption that they are sufficiently far from the region of interest so as to not affect the prediction (because cases with nearby ligands are filtered in an earlier stage),

2. had parameters assigned through a process requiring much intervention on a per case basis.

In order to improve upon this situation we have created a library of parameters for 11,837 small molecule ligands. Protonation states and tautomers for each ligand were identified using Schrodinger's epik. Parameters for bonded and non-bonded terms were assigned using the OPLS 2005 force field using macromodel. A static ligand "core" is identified and groups with rotable bonds attached to this core are identified. For these flexible groups, a rotamer library is constructed at a uniform resolution of 10 degrees to allow sampling of flexible ligands. Using these rotamer libraries it is possible sample conformations of flexible ligands by treating flexible groups in a fashion similar protein side chains. However due to lack of experimental conformations these rotamer libraries are not knowledge based as in the case of protein side chains.

The performance of a selection of the energy models implemented in PLOP was compared using a subset of these ligands found in CDK2 crystal structures. Minimized crystal structures starting from the crystal coordinates were consistenly closer to native conformations using the variable dielectric generalized Born solvation model than using a constant internal dielectric. However, there was no differentiation found between the newer and significantly more complex energy model, variously named optimized variable dielectric (OVD) and variable dielectric surface generalized Born 2.0 (vsgb2.0), which implements a number of terms to explicitly describe hydrogen bonding, $\pi - \pi$ interactions, and a number of other organic interactions.

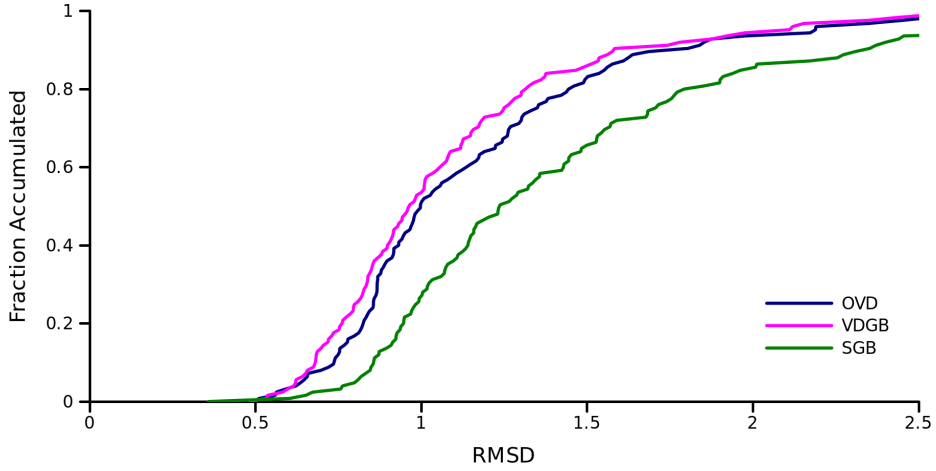


Figure 5.2: The fraction of minimized structures found within a given RMSD to native. The newer energy models, optimized variable dielectric (OVD or VSGB2.0) and variable dielectric surface generalized Born (VSGB) perform better than the original surface generalized Born model, however there is not significant differentiation between the two.

5.3 Knowledge Based Backbone Dihedral Penalty

For a number of challenging cases, we experimented with the use of a new addition to our energy model that penalizes loop conformations that are constructed with seldom-observed dipeptide dihedrals. The dipeptide rotamer frequency-based scoring term employed a greatly expanded dipeptide rotamer library (garnered from ~7500 high-quality PDB structures) that incorporated the frequency of each rotamer in this subset of the PDB. This information was used to penalize loop dipeptides whose combination of (ϕ, ψ) angles fall in an extremely unpopulated region of the five-dimensional dipeptide analogue to the well-known Ramachandran plot. The set of five angles for each dipeptide in the predicted loop, using a "sliding window" scheme, is compared against the new library to find the nearest dipeptide rotamer. Two criteria determine whether a penalty will be applied to the dipeptide:

1. if the Euclidean distance between the loop dipeptide and the nearest rotamer in the library is greater than a certain, empirically determined cutoff

2. if the total population of rotamers within a set radius of the loop dipeptide is below a

certain threshold The form of this penalty term, its implementation, and its successes in improving loop prediction in crystal structure and homology model environments will be discussed in detail in an upcoming publication.

This term was used in two situations:

1. for all of the predictions in inexact environments this is a substantially more challenging sampling and scoring problem, and the information contained in the dipeptide score can be expected to improve results systematically,

2. for a small subset of the predictions in the native environment where difficulties in the standard prediction approach were encountered.

To date, we have not found any cases where this term worsens results. However, more extensive tests are underway and will be presented in a subsequent publication. Dot (directed graphs) [Koutsofios *et al.*, 1991]

Bibliography

[Adams and Brantner, 2006] Christopher P Adams and Van V Brantner. Estimating the cost of new drug development: is it really $802 million? *Health Affairs*, 25(2):420–428, 2006.

[Afzelius *et al.*, 2007] Lovisa Afzelius, Catrin Hasselgren Arnby, Anders Broo, Lars Carlsson, Christine Isaksson, Ulrik Jurva, Britta Kjellander, Karin Kolmodin, Kristina Nilsson, Florian Raubacher, et al. State-of-the-art tools for computational site of metabolism predictions: comparative analysis, mechanistical insights, and future applications. *Drug metabolism reviews*, 39(1):61–86, 2007.

[Agresti *et al.*, 2010] Jeremy J Agresti, Eugene Antipov, Adam R Abate, Keunho Ahn, Amy C Rowat, Jean-Christophe Baret, Manuel Marquez, Alexander M Klibanov, Andrew D Griffiths, and David A Weitz. Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proceedings of the National Academy of Sciences*, 107(9):4004–4009, 2010.

[Bajorath and Sheriff, 1996] Jürgen Bajorath and Steven Sheriff. Comparison of an antibody model with an x-ray structure: The variable fragment of br96. *Proteins: Structure, Function, and Bioinformatics*, 24(2):152–157, 1996.

[Barreiro *et al.*, 2007] Gabriela Barreiro, Joseph T Kim, Cristiano RW Guimarães, Christopher M Bailey, Robert A Domaoal, Ligong Wang, Karen S Anderson, and William L Jorgensen. From docking false-positive to active anti-hiv agent. *Journal of medicinal chemistry*, 50(22):5324–5329, 2007.

[Bentley and Friedman, 1979] Jon Louis Bentley and Jerome H Friedman. Data structures for range searching. *ACM Computing Surveys (CSUR)*, 11(4):397–409, 1979.

[Berman *et al.*, 2007] Helen Berman, Kim Henrick, Haruki Nakamura, and John L Markley. The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data. *Nucleic acids research*, 35(suppl 1):D301–D303, 2007.

[Bixon and Lifson, 1967] M Bixon and S Lifson. Potential functions and conformations in cycloalkanes. *Tetrahedron*, 23(2):769–784, 1967.

[Bleicher *et al.*, 2003] Konrad H Bleicher, Hans-Joachim Böhm, Klaus Müller, and Alexander I Alanine. Hit and lead generation: beyond high-throughput screening. *Nature Reviews Drug Discovery*, 2(5):369–378, 2003.

[Braden *et al.*, 1996] Bradford C Braden, Barry A Fields, Xavier Ysern, William Dall'Acqua, Fernando A Goldbaum, Roberto J Poljak, and Roy A Mariuzza. Crystal structure of an fv–fv idiotope–anti-idiotope complex at 1.9 å resolution. *Journal of molecular biology*, 264(1):137–151, 1996.

[Bruccoleri and Karplus, 1985] Robert E Bruccoleri and Martin Karplus. Chain closure with bond angle variations. *Macromolecules*, 18(12):2767–2773, 1985.

[Bryngelson and Wolynes, 1987] Joseph D Bryngelson and Peter G Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences*, 84(21):7524–7528, 1987.

[Buckle *et al.*, 1994] Ashley M Buckle, Gideon Schreiber, and Alan R Fersht. Protein-protein recognition: Crystal structural analysis of a barnase-barstar complex at 2.0-. ang. resolution. *Biochemistry*, 33(30):8878–8889, 1994.

[Canutescu and Dunbrack, 2003] Adrian A Canutescu and Roland L Dunbrack. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science*, 12(5):963–972, 2003.

[Chaplin, 2013] Martin Chaplin. Protein folding and denaturation. `http://www.lsbu.ac.uk/water/protein2.html`, 2013.

[Chen *et al.*, 1992] Lorenzo H Chen, GL Kenyon, F Curtin, S Harayama, ME Bembenek, GHOLAMHOSSEIN Hajipour, and CP Whitman. 4-oxalocrotonate tautomerase, an enzyme composed of 62 amino acid residues per monomer. *Journal of Biological Chemistry*, 267(25):17716–17721, 1992.

[Chen *et al.*, 1997] Hao Chen, Marcel J de Groot, Nico PE Vermeulen, and Robert P Hanzlik. Oxidative n-dealkylation of p-cyclopropyl-n, n-dimethylaniline. a substituent effect on a radical-clock reaction rationalized by ab initio calculations on radical cation intermediates. *The Journal of Organic Chemistry*, 62(23):8227–8230, 1997.

[Chen *et al.*, 1998] Weiqiao Chen, Luke L Koenigs, Stella J Thompson, Raimund M Peter, Allan E Rettie, William F Trager, and Sidney D Nelson. Oxidation of acetaminophen to its toxic quinone imine and nontoxic catechol metabolites by baculovirus-expressed and purified human cytochromes p450 2e1 and 2a6. *Chemical research in toxicology*, 11(4):295–301, 1998.

[Chothia and Janin, 1975] Cyrus Chothia and Joël Janin. Principles of protein-protein recognition. *Nature*, 256(5520):705–708, 1975.

[Chothia, 1976] Cyrus Chothia. The nature of the accessible and buried surfaces in proteins. *Journal of molecular biology*, 105(1):1–12, 1976.

[Connolly, 1983] Michael L Connolly. Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16(5):548–558, 1983.

[Corsino *et al.*, 2009] Patrick Corsino, Nicole Horenstein, David Ostrov, Thomas Rowe, Mary Law, Amanda Barrett, George Aslanidi, W Douglas Cress, and Brian Law. A novel class of cyclin-dependent kinase inhibitors identified by molecular docking act through a unique mechanism. *Journal of Biological Chemistry*, 284(43):29945–29955, 2009.

[Cortés and Siméon, 2005] Juan Cortés and Thierry Siméon. Sampling-based motion planning under kinematic loop-closure constraints. In *Algorithmic Foundations of Robotics VI*, pages 75–90. Springer, 2005.

[Cruciani *et al.*, 2005] Gabriele Cruciani, Emanuele Carosati, Benoit De Boeck, Kantharaj Ethirajulu, Claire Mackie, Trevor Howe, and Riccardo Vianello. Metasite: understanding metabolism in human cytochromes from the perspective of the chemist. *Journal of medicinal chemistry*, 48(22):6970–6979, 2005.

[Cunningham and Wells, 1989] Brian C Cunningham and James A Wells. High-resolution epitope mapping of hgh-receptor interactions by alanine-scanning mutagenesis. *Science*, 244(4908):1081–1085, 1989.

[de Visser *et al.*, 2002] Sam P de Visser, François Ogliaro, Pankaz K Sharma, and Sason Shaik. What factors affect the regioselectivity of oxidation by cytochrome p450? a dft study of allylic hydroxylation and double bond epoxidation in a model reaction. *Journal of the American Chemical Society*, 124(39):11809–11826, 2002.

[Dembo and Steihaug, 1983] Ron S Dembo and Trond Steihaug. Truncated-newtono algorithms for large-scale unconstrained optimization. *Mathematical Programming*, 26(2):190–212, 1983.

[DiMasi *et al.*, 2003] Joseph A DiMasi, Ronald W Hansen, and Henry G Grabowski. The price of innovation: new estimates of drug development costs. *Journal of health economics*, 22(2):151–185, 2003.

[DiMasi, 2001] Joseph A DiMasi. Risks in new drug development: approval success rates for investigational drugs. *Clinical Pharmacology And Therapeutics St Louis*, 69(5):297–307, 2001.

[Duane *et al.*, 1987] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[Feifel *et al.*, 1993] N Feifel, K Kucher, L Fuchs, M Jedrychowski, E Schmidt, K-H Antonin, PR Bieck, and CH Gleiter. Role of cytochrome p4502d6 in the metabolism of brofaromine. *European journal of clinical pharmacology*, 45(3):265–269, 1993.

[Figueirido *et al.*, 1997] Francisco Figueirido, Ronald M Levy, Ruhong Zhou, and BJ Berne. Large scale simulation of macromolecules in solution: Combining the periodic fast mul-

tipole method with multiple time step integrators. *The Journal of chemical physics*, 106:9835, 1997.

[Fine *et al.*, 1986] RM Fine, H Wang, PS Shenkin, DL Yarmush, and C Levinthal. Predicting antibody hypervariable loop conformations ii: Minimization and molecular dynamics studies of mcpc603 from many randomly generated loop conformations. *Proteins: Structure, Function, and Bioinformatics*, 1(4):342–362, 1986.

[Fiser *et al.*, 2000] András Fiser, Richard Kinh Gian Do, and Andrej Šali. Modeling of loops in protein structures. *Protein science*, 9(9):1753–1773, 2000.

[Friesner *et al.*, 2004] Richard A Friesner, Jay L Banks, Robert B Murphy, Thomas A Halgren, Jasna J Klicic, Daniel T Mainz, Matthew P Repasky, Eric H Knoll, Mee Shelley, Jason K Perry, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of medicinal chemistry*, 47(7):1739–1749, 2004.

[Gallicchio and Levy, 2004] Emilio Gallicchio and Ronald M Levy. Agbnp: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *Journal of computational chemistry*, 25(4):479–499, 2004.

[Gallicchio *et al.*, 2002] Emilio Gallicchio, Linda Yu Zhang, and Ronald M Levy. The sgb/np hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *Journal of computational chemistry*, 23(5):517–529, 2002.

[Geppert *et al.*, 2010] Hanna Geppert, Martin Vogt, and Jurgen Bajorath. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *Journal of chemical information and modeling*, 50(2):205–216, 2010.

[Ghosh *et al.*, 1998] Avijit Ghosh, Chaya Sendrovic Rapp, and Richard A Friesner. Generalized born model based on a surface integral formulation. *The Journal of Physical Chemistry B*, 102(52):10983–10990, 1998.

[Go and Scheraga, 1970] Nobuhiro Go and Harold A Scheraga. Ring closure and local conformational deformations of chain molecules. *Macromolecules*, 3(2):178–187, 1970.

[Gohlke and Klebe, 2002] Holger Gohlke and Gerhard Klebe. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie International Edition*, 41(15):2644–2676, 2002.

[Goldfeld *et al.*, 2011] Dahlia A Goldfeld, Kai Zhu, Thijs Beuming, and Richard A Friesner. Successful prediction of the intra-and extracellular loops of four g-protein-coupled receptors. *Proceedings of the National Academy of Sciences*, 108(20):8275–8280, 2011.

[Goldfeld *et al.*, 2013] Dahlia A Goldfeld, Kai Zhu, Thijs Beuming, and Richard A Friesner. Loop prediction for a gpcr homology model: Algorithms and results. *Proteins: Structure, Function, and Bioinformatics*, 81(2):214–228, 2013.

[Gonzalez, 2005] Frank J Gonzalez. Role of cytochromes p450 in chemical toxicity and oxidative stress: studies with cyp2e1. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 569(1):101–110, 2005.

[Gram *et al.*, 1992] Hermann Gram, Lori-Anne Marconi, Carlos F Barbas, Thomas A Collet, Richard A Lerner, Angray S Kang, et al. In vitro selection and affinity maturation of antibodies from a naive combinatorial immunoglobulin library. *Proceedings of the National Academy of Sciences*, 89(8):3576–3580, 1992.

[Greer *et al.*, 1994] Jonathan Greer, John W Erickson, John J Baldwin, and Michael D Varney. Application of the three-dimensional structures of protein target molecules in structure-based drug design. *Journal of medicinal chemistry*, 37(8):1035–1054, 1994.

[Griffiths *et al.*, 1984] Gillian M Griffiths, Claudia Berek, Matti Kaartinen, and Cesar Milstein. Somatic mutation and the maturation of immune response to 2-phenyl oxazolone. 1984.

[Guengerich, 2001] F Peter Guengerich. Common and uncommon cytochrome p450 reactions related to metabolism and chemical toxicity. *Chemical research in toxicology*, 14(6):611–650, 2001.

[Guengerich, 2007] F Peter Guengerich. Cytochrome p450 and chemical toxicology. *Chemical research in toxicology*, 21(1):70–83, 2007.

[Guerois *et al.*, 2002] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, 320(2):369–387, 2002.

[Halgren *et al.*, 2004] Thomas A Halgren, Robert B Murphy, Richard A Friesner, Hege S Beard, Leah L Frye, W Thomas Pollard, and Jay L Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *Journal of medicinal chemistry*, 47(7):1750–1759, 2004.

[Hao *et al.*, 2010] Ge-Fei Hao, Guang-Fu Yang, and Chang-Guo Zhan. Computational mutation scanning and drug resistance mechanisms of hiv-1 protease inhibitors. *The Journal of Physical Chemistry B*, 114(29):9663–9676, 2010.

[Harris, 2008] Mark Harris. Cuda fluid simulation in nvidia physx. *Siggraph Asia*, pages 77–84, 2008.

[Hastings, 1970] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[Hawkins *et al.*, 1992] Robert E Hawkins, Stephen J Russell, and Greg Winter. Selection of phage antibodies by binding affinity: mimicking affinity maturation. *Journal of molecular biology*, 226(3):889–896, 1992.

[Hopkins and Groom, 2002] Andrew L Hopkins and Colin R Groom. The druggable genome. *Nature reviews Drug discovery*, 1(9):727–730, 2002.

[HTSRC, 2004] HTSRC. The rockefeller university high-throughput screening resource center. `http://www.rockefeller.edu/htsrc/`, 2004.

[Hu *et al.*, 2000] Zengjian Hu, Buyong Ma, Haim Wolfson, and Ruth Nussinov. Conservation of polar residues as hot spots at protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 39(4):331–342, 2000.

[Humphrey *et al.*, 1996] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[Inokuma *et al.*, 2013] Yasuhide Inokuma, Shota Yoshioka, Junko Ariyoshi, Tatsuhiko Arai, Yuki Hitora, Kentaro Takada, Shigeki Matsunaga, Kari Rissanen, and Makoto Fujita. X-ray analysis on the nanogram to microgram scale using porous complexes. *Nature*, 495(7442):461–466, 2013.

[Irwin and Shoichet, 2005] John J Irwin and Brian K Shoichet. Zinc-a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

[Jacobson *et al.*, 2002a] Matthew P Jacobson, Richard A Friesner, Zhexin Xiang, and Barry Honig. On the role of the crystal environment in determining protein side-chain conformations. *Journal of molecular biology*, 320(3):597–608, 2002.

[Jacobson *et al.*, 2002b] Matthew P Jacobson, George A Kaminski, Richard A Friesner, and Chaya S Rapp. Force field validation using protein side chain prediction. *The Journal of Physical Chemistry B*, 106(44):11673–11680, 2002.

[Jacobson *et al.*, 2004] Matthew P Jacobson, David L Pincus, Chaya S Rapp, Tyler JF Day, Barry Honig, David E Shaw, and Richard A Friesner. A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics*, 55(2):351–367, 2004.

[Jones *et al.*, 1999] Susan Jones, Paul van Heyningen, Helen M Berman, and Janet M Thornton. Protein-dna interactions: a structural analysis. *Journal of molecular biology*, 287(5):877–896, 1999.

[Jorgensen and Tirado-Rives, 1988] William L Jorgensen and Julian Tirado-Rives. The opls [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666, 1988.

[Jorgensen *et al.*, 1996] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.

[Jorgensen, 2004] William L Jorgensen. The many roles of computation in drug discovery. *Science*, 303(5665):1813–1818, 2004.

[Jorgensen, 2009] William L Jorgensen. Efficient drug lead discovery and optimization. *Accounts of chemical research*, 42(6):724–733, 2009.

[Kaminski *et al.*, 1994] George Kaminski, Erin M Duffy, Tooru Matsui, and William L Jorgensen. Free energies of hydration and pure liquid properties of hydrocarbons from the opls all-atom model. *The Journal of Physical Chemistry*, 98(49):13077–13082, 1994.

[Kerns and Di, 2008] Edward Kerns and Li Di. *Drug-like properties: concepts, structure design and methods: from ADME to toxicity optimization*. Academic Press, 2008.

[Keserű and Makara, 2006] György M Keserű and Gergely M Makara. Hit discovery and hit-to-lead approaches. *Drug discovery today*, 11(15):741–748, 2006.

[Kivisto *et al.*, 1995] Kari T Kivisto, Heyo K Kroemer, and Michel Eichelbaum. The role of human cytochrome p450 enzymes in the metabolism of anticancer agents: implications for drug interactions. *British journal of clinical pharmacology*, 40(6):523–530, 1995.

[Klapper, 1977] Michael H Klapper. The independent distribution of amino acid near neighbor pairs into polypeptides. *Biochemical and biophysical research communications*, 78(3):1018–1024, 1977.

[Kolodny *et al.*, 2005] Rachel Kolodny, Leonidas Guibas, Michael Levitt, and Patrice Koehl. Inverse kinematics in biology: the protein loop closure problem. *The International Journal of Robotics Research*, 24(2-3):151–163, 2005.

[Kortemme and Baker, 2002] Tanja Kortemme and David Baker. A simple physical model for binding energy hot spots in protein–protein complexes. *Proceedings of the National Academy of Sciences*, 99(22):14116–14121, 2002.

[Kortemme *et al.*, 2004] Tanja Kortemme, David E Kim, and David Baker. Computational alanine scanning of protein-protein interfaces. *Science Signaling*, 2004(219):pl2, 2004.

[Koutsofios *et al.*, 1991] Eleftherios Koutsofios, Stephen North, et al. Drawing graphs with dot. Technical report, Technical Report 910904-59113-08TM, AT&T Bell Laboratories, Murray Hill, NJ, 1991.

[Kuntz *et al.*, 1982] Irwin D Kuntz, Jeffrey M Blaney, Stuart J Oatley, Robert Langridge, and Thomas E Ferrin. A geometric approach to macromolecule-ligand interactions. *Journal of molecular biology*, 161(2):269–288, 1982.

[Lee *et al.*, 1999] Jooyoung Lee, Adam Liwo, and Harold A Scheraga. Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: Application to the 10-55 fragment of staphylococcal protein a and to apo calbindin d9k. *Proceedings of the National Academy of Sciences*, 96(5):2025–2030, 1999.

[Leopold *et al.*, 1992] Peter E Leopold, Mauricio Montal, and José N Onuchic. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences*, 89(18):8721–8725, 1992.

[Leung and White, 1989] Hareton KN Leung and Lee White. Insights into regression testing [software testing]. In *Software Maintenance, 1989., Proceedings., Conference on*, pages 60–69. IEEE, 1989.

[Levinthal, 1966] Cyrus Levinthal. *Molecular model-building by computer*. WH Freeman and Company, 1966.

[Levitt and Lifson, 1969] Michael Levitt and Shneior Lifson. Refinement of protein conformations using a macromolecular energy minimization procedure. *Journal of molecular biology*, 46(2):269–279, 1969.

[Li and Scheraga, 1987] Zhenqin Li and Harold A Scheraga. Monte carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences*, 84(19):6611–6615, 1987.

[Li *et al.*, 2004] Xianchun Li, Jerome Baudry, May R Berenbaum, and Mary A Schuler. Structural and functional divergence of insect cyp6b proteins: From specialist to generalist cytochrome p450. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2939–2944, 2004.

[Li *et al.*, 2007] X. Li, M.P. Jacobson, K. Zhu, S. Zhao, and R.A. Friesner. Assignment of polar states for protein amino acid residues using an interaction cluster decomposition algorithm and its application to high resolution protein structure modeling. *Proteins: Structure, Function, and Bioinformatics*, 66(4):824–837, 2007.

[Li *et al.*, 2011a] Jianing Li, Robert Abel, Kai Zhu, Yixiang Cao, Suwen Zhao, and Richard A Friesner. The vsgb 2.0 model: a next generation energy model for high resolution protein structure modeling. *Proteins: Structure, Function, and Bioinformatics*, 79(10):2794–2812, 2011.

[Li *et al.*, 2011b] Jianing Li, Severin T Schneebeli, Joseph Bylund, Ramy Farid, and Richard A Friesner. Idsite: an accurate approach to predict p450-mediated drug metabolism. *Journal of chemical theory and computation*, 7(11):3829–3845, 2011.

[Li, 2001] Albert P Li. Screening for human adme/tox drug properties in drug discovery. *Drug discovery today*, 6(7):357–366, 2001.

[Lichtarge *et al.*, 1996] Olivier Lichtarge, Henry R Bourne, and Fred E Cohen. An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology*, 257(2):342–358, 1996.

[Linnet and Olesen, 1997] Kristian Linnet and Ole V Olesen. Metabolism of clozapine by cdna-expressed human cytochrome p450 enzymes. *Drug metabolism and disposition*, 25(12):1379–1382, 1997.

[Lipinski *et al.*, 1997] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1):3–25, 1997.

[MacKerell *et al.*,] Alexander D MacKerell, Bernard Brooks, Charles L Brooks, Lennart Nilsson, Benoit Roux, Youngdo Won, and Martin Karplus. Charmm: The energy function and its parameterization. *Encyclopedia of computational chemistry*.

[Massova and Kollman, 1999] Irina Massova and Peter A Kollman. Computational alanine scanning to probe protein-protein interactions: a novel approach to evaluate binding free energies. *Journal of the American Chemical Society*, 121(36):8133–8143, 1999.

[Metropolis *et al.*, 1953] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.

[Miller *et al.*, 2013] Edward B Miller, Colleen S Murrett, Kai Zhu, Suwen Zhao, Dahlia A Goldfeld, Joseph H Bylund, and Richard A Friesner. Prediction of long loops with embedded secondary structure using the protein local optimization program. *Journal of Chemical Theory and Computation*, 9(3):1846–1864, 2013.

[Moult and James, 1986] James Moult and MNG James. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins: Structure, Function, and Bioinformatics*, 1(2):146–163, 1986.

[MSSR, 2006] MSSR. Molecular screening shared resource (mssr). `http://www.mssr.ucla.edu/`, 2006.

[Nicholls and Honig, 1991] Anthony Nicholls and Barry Honig. A rapid finite difference algorithm, utilizing successive over-relaxation to solve the poisson–boltzmann equation. *Journal of computational chemistry*, 12(4):435–445, 1991.

[Paine *et al.*, 2003] Mark JI Paine, Lesley A McLaughlin, Jack U Flanagan, Carol A Kemp, Michael J Sutcliffe, Gordon CK Roberts, and C Roland Wolf. Residues glutamate 216 and aspartate 301 are key determinants of substrate specificity and product regioselectivity in cytochrome p450 2d6. *Journal of Biological Chemistry*, 278(6):4021–4027, 2003.

[Palmer and Scheraga, 1991] Kathleen A Palmer and Harold A Scheraga. Standard-geometry chains fitted to x-ray derived structures: Validation of the rigid-geometry ap-

proximation. i. chain closure through a limited search of loop conformations. *Journal of computational chemistry*, 12(4):505–526, 1991.

[Paul *et al.*, 2010] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature reviews Drug discovery*, 9(3):203–214, 2010.

[Ponder and Richards, 1987] Jay W Ponder and Frederic M Richards. An efficient newton-like method for molecular mechanics energy minimization of large molecules. *Journal of Computational Chemistry*, 8(7):1016–1024, 1987.

[Pons *et al.*, 1999] Jaume Pons, Arvind Rajpal, and Jack F Kirsch. Energetic analysis of an antigen/antibody interface: Alanine scanning mutagenesis and double mutant cycles on the hyhel-10/lysozyme interaction. *Protein science*, 8(5):958–968, 1999.

[POVRAY 3.6, 2004] Persistence of Vision Pty. Ltd. POVRAY 3.6. Persistence of vision raytracer. `http://www.povray.org/`, 2004. Version 3.6.

[Proudfoot, 2002] John R Proudfoot. Drugs, leads, and drug-likeness: an analysis of some recently launched drugs. *Bioorganic & medicinal chemistry letters*, 12(12):1647–1650, 2002.

[Qiu *et al.*, 1997] Di Qiu, Peter S Shenkin, Frank P Hollinger, and W Clark Still. The gb/sa continuum model for solvation. a fast analytical method for the calculation of approximate born radii. *The Journal of Physical Chemistry A*, 101(16):3005–3014, 1997.

[Rakhmanov *et al.*, 1994] EA Rakhmanov, EB Saff, and YM Zhou. Minimal discrete energy on the sphere. *Math. Res. Lett*, 1(6):647–662, 1994.

[Rapp and Friesner, 1999] Chaya Sendrovic Rapp and Richard A Friesner. Prediction of loop geometries using a generalized born model of solvation effects. *Proteins: Structure, Function, and Bioinformatics*, 35(2):173–183, 1999.

[Richards, 1977] F M Richards. Areas, volumes, packing, and protein structure. *Annual Review of Biophysics and Bioengineering*, 6(1):151–176, 1977. PMID: 326146.

[Richmond, 1984] Timothy J Richmond. Solvent accessible surface area and excluded volume in proteins: Analytical equations for overlapping spheres and implications for the hydrophobic effect. *Journal of molecular biology*, 178(1):63–89, 1984.

[Rose *et al.*, 1985] George D Rose, Ari R Geselowitz, Glenn J Lesser, Richard H Lee, and Micheal H Zehfus. Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716):834–838, 1985.

[Roux and Simonson, 1999] Benoıt Roux and Thomas Simonson. Implicit solvent models. *Biophysical Chemistry*, 78(1):1–20, 1999.

[Rowland *et al.*, 2006] Paul Rowland, Frank E Blaney, Martin G Smyth, Jo J Jones, Vaughan R Leydon, Amanda K Oxbrow, Ceri J Lewis, Mike G Tennant, Sandeep Modi, Drake S Eggleston, et al. Crystal structure of human cytochrome p450 2d6. *Journal of Biological Chemistry*, 281(11):7614–7622, 2006.

[Saff and Kuijlaars, 1997] Edward B Saff and A BJ Kuijlaars. Distributing many points on a sphere. *The Mathematical Intelligencer*, 19(1):5–11, 1997.

[Sauer-Eriksson *et al.*, 1995] A Elisabeth Sauer-Eriksson, Gerard J Kleywegt, Mathias Uhlén, and T Alwyn Jones. Crystal structure of the c2 fragment of streptococcal protein g in complex with the fc domain of human igg. *Structure*, 3(3):265–278, 1995.

[Scannell *et al.*, 2012] Jack W Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. Diagnosing the decline in pharmaceutical r&d efficiency. *Nature reviews Drug discovery*, 11(3):191–200, 2012.

[Schaefer and Karplus, 1996] Michael Schaefer and Martin Karplus. A comprehensive analytical treatment of continuum electrostatics. *The Journal of Physical Chemistry*, 100(5):1578–1599, 1996.

[Schlick, 2010] Tamar Schlick. *Molecular modeling and simulation: an interdisciplinary guide*, volume 21. Springer, 2010.

[Scott *et al.*, 2004] Emily E Scott, Mark A White, You Ai He, Eric F Johnson, C David Stout, and James R Halpert. Structure of mammalian cytochrome p450 2b4 complexed

with 4-(4-chlorophenyl) imidazole at 1.9-å resolution insight into the range of p450 conformations and the coordination of redox partner binding. *Journal of Biological Chemistry*, 279(26):27294–27301, 2004.

[Sellers *et al.*, 2008] Benjamin D Sellers, Kai Zhu, Suwen Zhao, Richard A Friesner, and Matthew P Jacobson. Toward better refinement of comparative models: predicting loops in inexact environments. *Proteins: Structure, Function, and Bioinformatics*, 72(3):959–971, 2008.

[Shaik *et al.*, 2005] Sason Shaik, Devesh Kumar, Samuël P de Visser, Ahmet Altun, and Walter Thiel. Theoretical perspective on the structure and mechanism of cytochrome p450 enzymes. *Chemical reviews*, 105(6):2279–2328, 2005.

[Sharp, 2001] Kim Sharp. Entropyenthalpy compensation: Fact or artifact? *Protein Science*, 10(3):661–667, 2001.

[Shenkin *et al.*, 1987] Peter S Shenkin, David L Yarmush, Richard M Fine, Huajun Wang, and Cyrus Levinthal. Predicting antibody hypervariable loop conformation. i. ensembles of random conformations for ringlike structures. *Biopolymers*, 26(12):2053–2085, 1987.

[Sheridan *et al.*, 2007] Robert P Sheridan, Kenneth R Korzekwa, Rhonda A Torres, and Matthew J Walker. Empirical regioselectivity models for human cytochromes p450 3a4, 2d6, and 2c9. *Journal of medicinal chemistry*, 50(14):3173–3184, 2007.

[Shoichet, 2004] Brian K Shoichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, 2004.

[Shrake and Rupley, 1973] A Shrake and JA Rupley. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of molecular biology*, 79(2):351–371, 1973.

[Singh *et al.*, 2003] Suresh B Singh, Lucy Q Shen, Matthew J Walker, and Robert P Sheridan. A model for predicting likely sites of cyp3a4-mediated metabolism on drug-like molecules. *Journal of medicinal chemistry*, 46(8):1330–1336, 2003.

[Smith, 1985] George P Smith. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, 228(4705):1315–1317, 1985.

[Still *et al.*, 1990] W Clark Still, Anna Tempczyk, Ronald C Hawley, and Thomas Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society*, 112(16):6127–6129, 1990.

[Thorn and Bogan, 2001] Kurt S Thorn and Andrew A Bogan. Asedb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17(3):284–285, 2001.

[Torrie and Valleau, 1977] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

[Tsai *et al.*, 1999] Chung-Jung Tsai, Sandeep Kumar, Buyong Ma, and Ruth Nussinov. Folding funnels, binding funnels, and protein function. *Protein Science*, 8(6):1181–1190, 1999.

[Turk, 1989] Greg Turk. *Interactive collision detection for molecular graphics*. PhD thesis, The University of North Carolina, 1989.

[Wang and Chen, 1991] L-CT Wang and Chih Cheng Chen. A combined optimization method for solving the inverse kinematics problems of mechanical manipulators. *Robotics and Automation, IEEE Transactions on*, 7(4):489–499, 1991.

[Wang *et al.*, 2007] Yonghua Wang, Yan Li, and Bin Wang. Stochastic simulations of the cytochrome p450 catalytic cycle. *The Journal of Physical Chemistry B*, 111(16):4251–4260, 2007.

[Wedemeyer and Scheraga, 1999] William J Wedemeyer and Harold A Scheraga. Exact analytical loop closure in proteins using polynomial equations. *Journal of Computational Chemistry*, 20(8):819–844, 1999.

[Weide *et al.*, 1995] Bruce W Weide, Wayne D Heym, and Joseph E Hollingsworth. Reverse engineering of legacy code exposed. In *Proceedings of the 17th international conference on Software engineering*, pages 327–331. ACM, 1995.

[Weiner *et al.*, 1984] Scott J Weiner, Peter A Kollman, David A Case, U Chandra Singh, Caterina Ghio, Guliano Alagona, Salvatore Profeta, and Paul Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, 1984.

[Williams *et al.*, 2004] J Andrew Williams, Ruth Hyland, Barry C Jones, Dennis A Smith, Susan Hurst, Theunis C Goosen, Vincent Peterkin, Jeffrey R Koup, and Simon E Ball. Drug-drug interactions for udp-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (auci/auc) ratios. *Drug Metabolism and Disposition*, 32(11):1201–1208, 2004.

[Wong *et al.*, 1997] W Eric Wong, Joseph R Horgan, Saul London, and Hira Agrawal. A study of effective regression testing in practice. In *PROCEEDINGS The Eighth International Symposium On Software Reliability Engineering*, pages 264–274. IEEE, 1997.

[Wu and Dean, 1996] Sheng-Jiun Wu and Donald H Dean. Functional significance of loops in the receptor binding domain of¡ i¿ bacillus thuringiensis¡/i¿ cryiiia δ-endotoxin. *Journal of molecular biology*, 255(4):628–640, 1996.

[Xiang and Honig, 2001] Zhexin Xiang and Barry Honig. Extending the accuracy limits of prediction for side-chain conformations. *Journal of molecular biology*, 311(2):421–430, 2001.

[Zhang *et al.*, 2001] Linda Yu Zhang, Emilio Gallicchio, Richard A Friesner, and Ronald M Levy. Solvent models for protein–ligand binding: Comparison of implicit solvent poisson and surface generalized born models with explicit solvent simulations. *Journal of Computational Chemistry*, 22(6):591–607, 2001.

[Zhao *et al.*, 2011] Suwen Zhao, Kai Zhu, Jianing Li, and Richard A Friesner. Progress in super long loop prediction. *Proteins: Structure, Function, and Bioinformatics*, 79(10):2920–2935, 2011.

[Zhou, 1995] Yanmu Zhou. Arrangements of points on the sphere. 1995.

[Zhou, 2003] Ruhong Zhou. Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins: Structure, Function, and Bioinformatics*, 53(2):148–161, 2003.

[Zhu *et al.*, 2006] Kai Zhu, David L Pincus, Suwen Zhao, and Richard A Friesner. Long loop prediction using the protein local optimization program. *Proteins: Structure, Function, and Bioinformatics*, 65(2):438–452, 2006.

[Zhu *et al.*, 2007] Kai Zhu, Michael R Shirts, and Richard A Friesner. Improved methods for side chain and loop predictions via the protein local optimization program: Variable dielectric model for implicitly improving the treatment of polarization effects. *Journal of Chemical Theory and Computation*, 3(6):2108–2119, 2007.