Table of Contents

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.

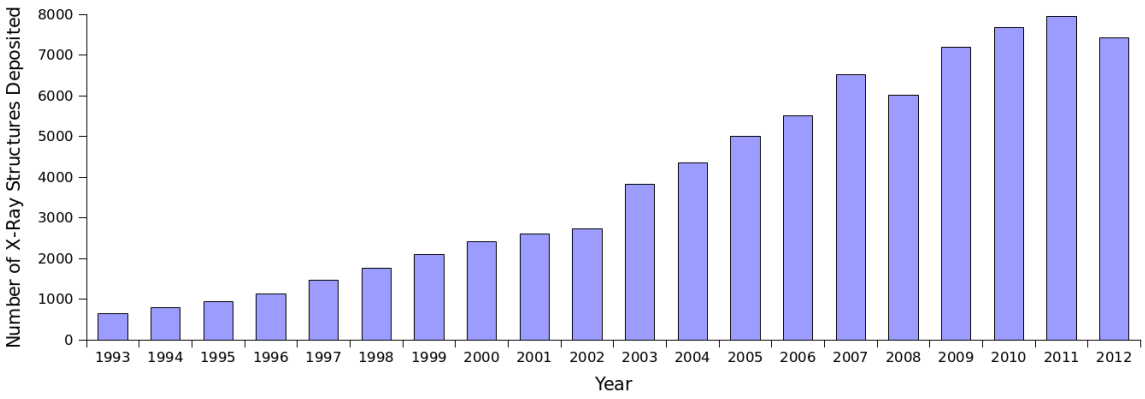


Figure 1.1: The rate at which new structures is deposited into the PDB over the last two decades. Due to a variety of improvements in the field of crystallography this rate has been steadily increasing.

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].

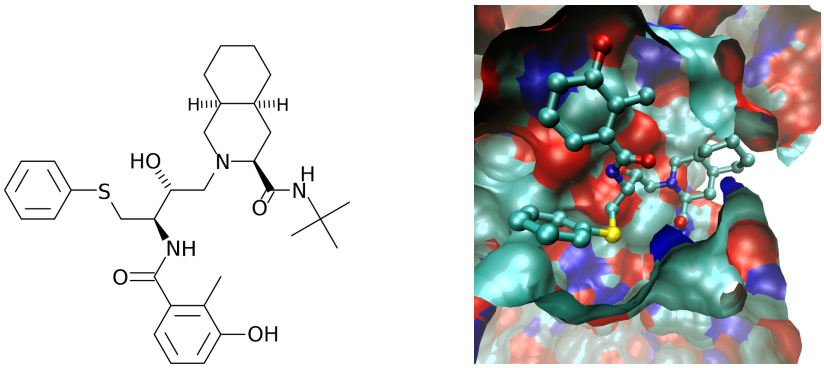
Of the total proteome only 30,000 are regulated by small molecule binding, making them reasonable targets of drug action. A large number of these possible drug targets are not implicated in any disease, due to this and a number of other factors, estimates of the total number of these proteins which are 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

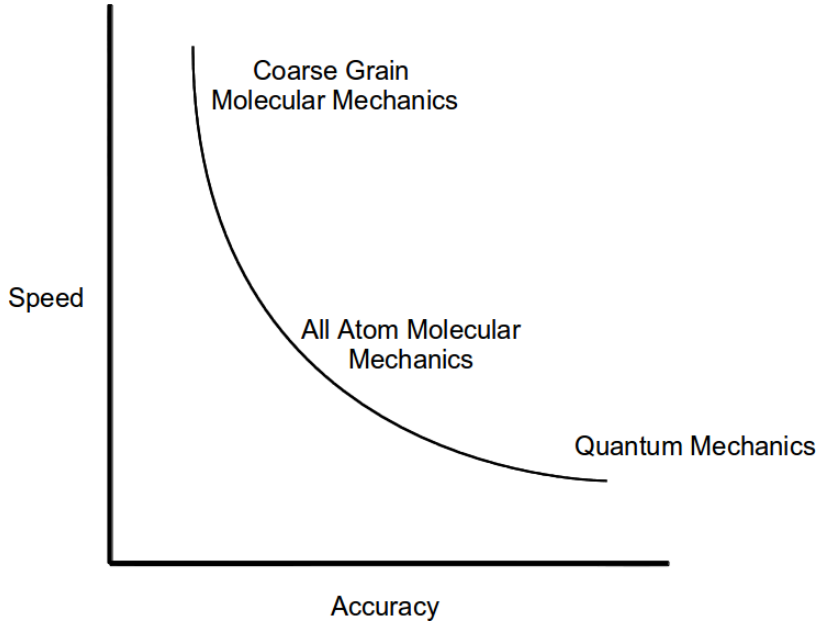


Figure 1.3: To an extent it is always possible to either increase accuracy or decrease running time, or the cost of an experiment. New scientific methods should allow one to increase accuracy while not spending additional time.

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.

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:

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].

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

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

Prediction of P450 Sites of Metabolism

2.1 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]. 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

$$\mathrm{RH} + \mathrm{O_2} + \mathrm{NADPH} + \mathrm{H^+} > \mathrm{ROH} + \mathrm{H_2O} + \mathrm{NADP^+} \tag{2.1}$$

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

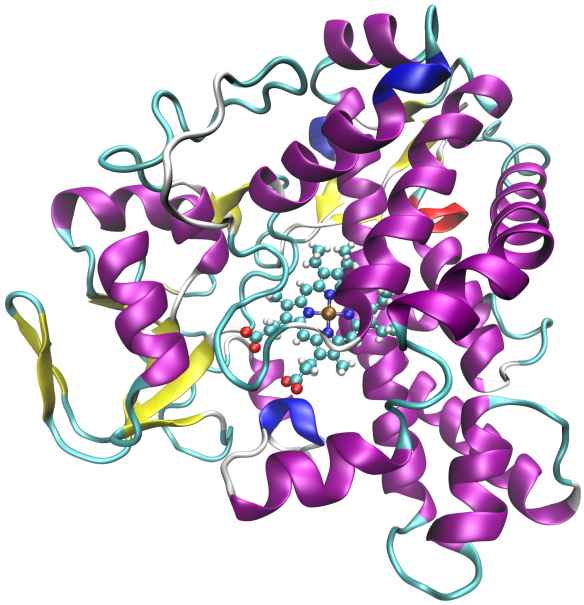


Figure 2.1: The structure of cytochrome P450, taken from PDBid 1JFB, shown in cartoon representation. The bonded heme group, shown as ball and stick model, is visible in the center. The brown iron atom is chelated by four deep blue nitrogen atoms.

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.

2.2 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) 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

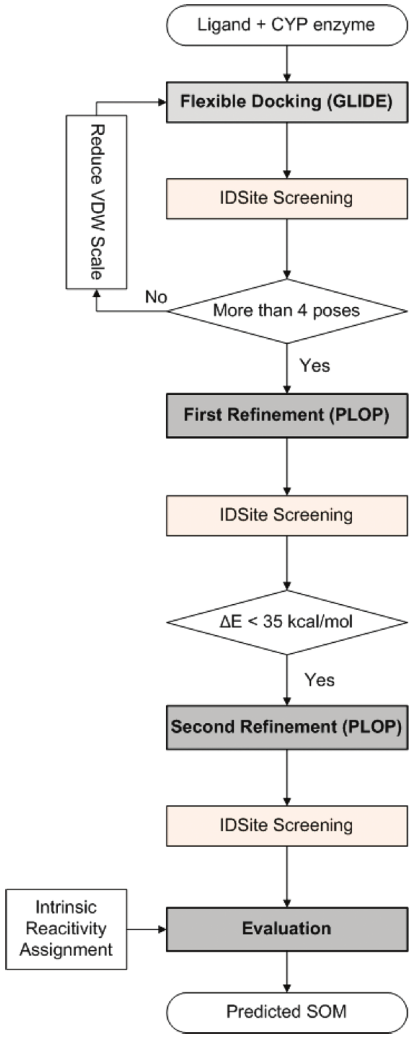


Figure 2.2: An overview of the entire IDSite procedure.

of the energy of the refined conformations and the intrinsic reactivity of the site. [Li *et al.*, 2011]

2.2.1 Docking

Starting from the ligand and the protein receptor structures, IDSite carries out flexible ligand docking with Glide. The flexible ligand docking protocol generates a large number of ligand conformations that are then docked into the rigid receptor. The first step in Glide docking is to define the binding box and calculate the receptor grid. As in Glide, in IDSite the binding site is defined as a box centered at the center of selected residues or

a ligand (if the structure contains a ligand). Because we start from the apo structure of CYP2D6 (PDBID: 2F9Q. See below for details about the protein preparation), the center of the binding box is selected as the centroid of the residues Glu216, Asp301, Thr309, and Phe483. The box dimension on each side is set to 10 angstroms for the inner box and 20 angstroms for the outer box. After the grid generation, IDSite samples the conformations of freely rotatable bonds and rings with Glide Standard Precision (SP). In order to increase sampling, IDSite uses reduced Van der Waals (VDW) radii and skips the default filtering with a rough score within Glide (also referred to as expanded sampling). Similar poses are clustered according to their RMSD (cutoff 2.0 angstroms). Finally, a post-docking minimization is performed and the top 60 minimized poses according to the Glide SP score are retained. These poses are then screened to remove the poses with obvious steric clashes, with too many atoms outside the inner binding box, or without atoms close to the heme iron (Table 3.1). The remaining poses are then passed to the first refinement stage. IDSite uses reduced VDW radii for nonpolar atoms both in the protein receptor and the ligand, so that slight steric clashes are tolerated during the docking stage. For the protein receptor the VDW scaling factor is fixed at 0.40, while for the ligand, the scaling factor starting from 0.80 is adaptively adjusted until at least 4 valid poses are found. With highly flexible ligands and relatively high scaling factors, Glide often finds only a handful of valid poses, and even fewer survive after IDSite screening. However, if the scaling factor is set too low, the docked poses may contain too many serious steric clashes, which can cause problems in the subsequent minimization. If IDSite fails to find enough valid poses, the scaling factor is adjusted and the number of poses to pass the initial docking phase in Glide is increased accordingly to augment sampling. Since a typical CYP2D6 substrate forms a highly conserved salt bridge with either Glu216 or Asp301,125 IDSite employs this conserved interaction to reduce the sampling cost of the CYP2D6-docking in the following way: IDSite adds a positional constraint to ensure that the generated poses fulfill at least part of the preferred conserved interactions. The positional constraint defines a spherical region in the receptor that is within 4.0 angstroms of the center of the Glu216, Asp301, and Ser304 residues (Figure 3.2). It is required that during docking and post- docking minimization each pose should maintain at least one hydrogen-bond donor inside the spherical region. If the

ligand contains other hydrogen-bond donors except for the basic nitrogen, the constrained docking is likely to generate poses that form hydrogen bonds instead of the salt bridge to Glu216 or Asp301. However, IDSite is able to distinguish these poses and filter them via an additional salt bridge filter in the pose screening (Table 3.1), so that only the poses with a stable salt bridge are allowed to pass to the refinement stage.

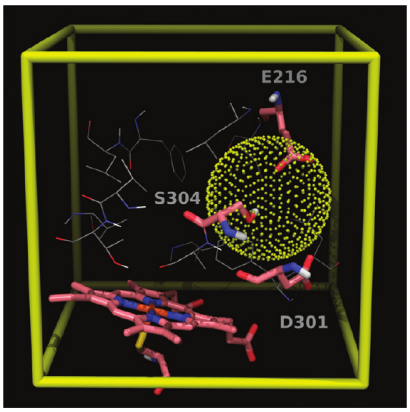


Figure 2.3: An overview of the entire IDSite procedure.

2.2.2 Monte Carlo Minimization Refinement

The refinement procedure used by IDSITE, makes use of a Monte Carlo Minimization procedure. Since an emphasis was placed on efficient sampling of low energy conformations, as only these 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). This IDSITE Monte Carlo Minimization refinement procedure incorportates three different types of steps: side chain motions, rigid body transformations, and hybrid Monte Carlo simulations. For each Monte Carlo step one of these three types of motions is selected according to the weighted probabilities 14:7:1 (side-chain:rigid-body:hybrid-Monte-Carlo). The standard Metropolis acceptance criteria (equation 1.1) is to the structure during each outer loop step, using a temperature that depends on the step type. 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.

1. Side chain moves. Several types of side chain motions were implemented in PLOP. In all cases, they are defined in a such a way that they can be applied to both ligands and proteins. The same atomic overlap screening function implemented with the rigid body Monte Carlo was implemented with the side chain torsional moves. a. Random torsion angle moves: The first type of move that was implemented is random movement of torsional chi angles. For small torsion moves, a random perturbation of the angle of +/- X is made, where X is a random number with user defined magnitude. For large torsion moves, for each torsion angle that is changed, a random angle is selected in the form 60*Y +/- X, where Y = 1 through 5, and X is the same random number for the small torsion moves. The large move was introduced since positions at the top of rotamer barriers are relatively unlikely to be selected, and efficiency thus can be improved by focusing on the more probable moves. The ratio of small to large torsion moves can be used-adjusted, as can the ratio of probabilities of changing all the torsions in a randomly selected side chain versus changing only one single (randomly selected) torsion among all the free torsions in the simulation can be set as a user-defined parameter. Rotamer side chain moves: A second type of torsional samples implemented is random selection of a new rotamer state for the entire side chain, plus an optional user defined small noise term for each torsion in the rotamer state. A database of protein rotamer states obtained from crystallographic data are already a part of PLOP [Xiang and Honig, 2001] Rotamer libraries for ligands are generated by examining all possible side chain conformations at 10 degree resolution and screening this set for steric clashes. A Monte Carlo move in this case represents a choice of a new torsional rotamer state for the entire side chain. Monte Carlo moves based on torsional states cannot lead to correct equilibrium distributions, as transitions from non-rotamer states to rotamer states are defined, but not reverse transitions, upsetting detailed balance. However, a pretabulated rotamer state is more likely to be low energy than a randomly generated torsional state, and thus allows for more diverse conformational searching. Correlated torsional moves: Most torsional rearrangements of the side chains in the core of proteins are highly correlated because of the density. In order to attempt to include correlated torsional motion, at each step

we examine the distance between all pairs of beta carbons in the ligands that are free to move. At each step, for the set of side chains that are free to move, clusters where beta carbons are all mutually within a user-specified distance are identified. This process takes a trivial amount of time compared to an energy evaluation, so does not slow the simulation at all. Then, with user specified probabilities, clusters of different sizes are selected for the torsional moves, either with random side chain moves, or rotamer selection moves. By selecting only clusters where all residues are mutual neighbors, detailed balanced is observed for simulations where accurate equilibrium sampling is desired. By varying the dihedral angles of the rotatable bonds, IDSite uses side chain MC moves in PLOP to sample the selected side-chain conformations of the protein and of the ligand. Up to three close residues (C beta distance within 6 angstroms) are allowed to rotate collectively, but the moves of the protein residues and those of the ligand are separated. In each attempted movement, the conformations of the selected side chains (from the protein/ligand) are either changed by random perturbations or assigned by the randomly selected rotamers from a library. For an attempt with a random perturbation, the displacement of each dihedral angle is the sum of a large rotation (N times 60 degrees with N as a random integer between 0 and 5) and a random perturbation from 0 to 30 degrees. For a rotamer library attempt, a side-chain conformation is updated with a random rotamer from a high resolution side-chain library for protein residues [Xiang and Honig, 2001], and from a homogeneous library at 10 degree resolution for the ligand. If a structure with tolerable overlaps is generated in an attempt, it is minimized and sent to subsequent stages for judgment of acceptance. Each side-chain move takes less than 15 seconds and is the fastest among all the three move types.

For side chain Monte Carlo, a steric screen with an overlap factor of 0.6 was used. Rotamer torsional moves were selected 75% of the time, with half of the remaining being of random torsions, and the other half random perturbations of all torsions within the randomly selected side chains. Clusters of size 1 (i.e. single side chains), size 2 and size three were selected in equal proportion, and all side chains in the cluster were perturbed with the selected torsion move. A mutual beta carbon distance of 6

Angstroms was used for the clustering size. Small torsion perturbations made +/- 60 degress from the current dihedral angle, and were performed 5% of the time; Large periodic moves were performed 95% of the time. Only outer steps were performed, and each side chain Monte Carlo series consisted in only one move. Minimization was performed after the single step, and acceptance was performed at 1 K.

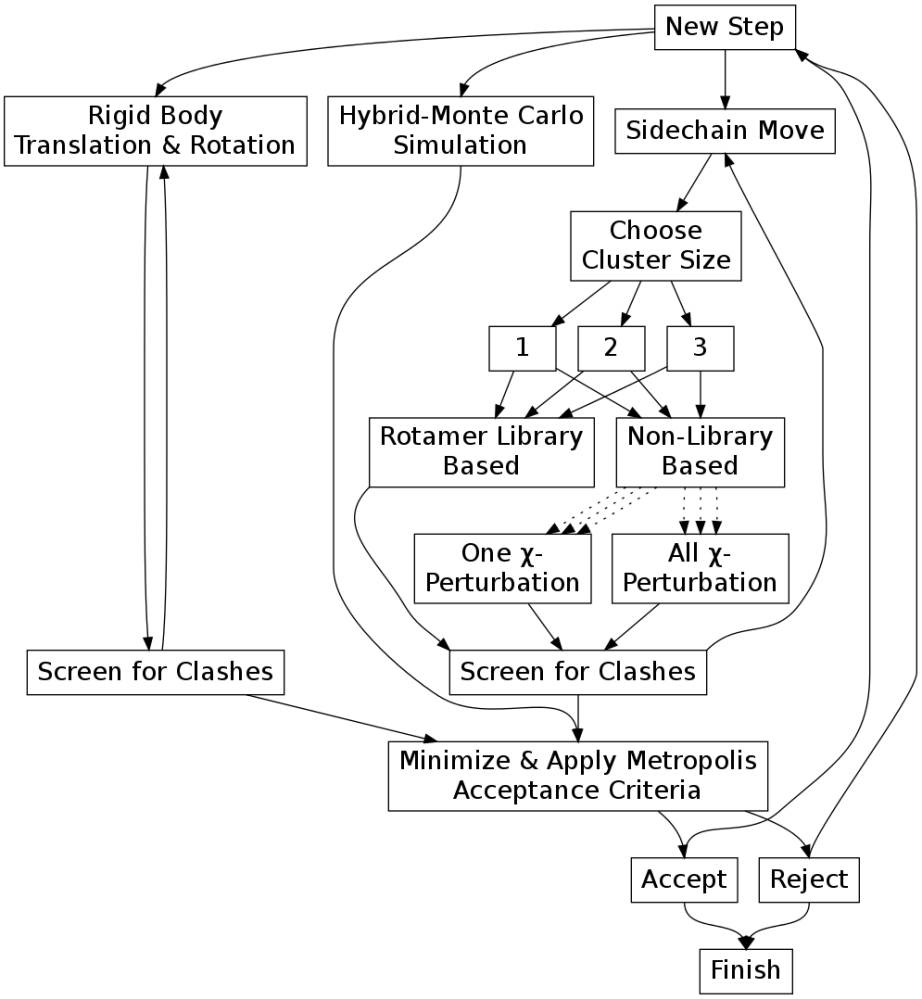


Figure 2.4: An outline of the Monte Carlo minimization refinement stages in PLOP.

2. Rigid motion moves. Rigid body translation and rotation were also implemented for noncovalently linked moieties, such as ligands. Random rotations and translations were coupled together, allowing for more concerted movement. Rigid body move implemented in PLOP can optionally include a screening step, where atomic Lennard-

Jones overlaps that would lead to energies much higher than would be observed in any conceivably long equilibrium simulation are rejected without further evaluation. A ratio of 0.7 between the distance between the two atoms and the sum of the Lennard-Jones radii of the two atoms yields energies on the order of 10's of thousands of kcal/mol, and is thus reasonable to maintain equilibrium sampling in a Monte Carlo simulation. Translations were implemented in a random direction, with a user-defined magnitude. Rotations were implemented by picking a random quarternion (a random angle around a random axis, through the geometric center of the rigid group) with a user specified maximum random angle centered around either the current angle, or 180 from the current angle, in the case of a flip. Multiple time scale Monte Carlo sampling was also implemented with rigid body moves, with short range and long range interactions defined as above. In addition, an option to compute the inner Monte Carlo loops with reduced Lennard-Jones radii were also implemented, to increase the ability to escape from tight spacial bottlenecks. In this case, the long time step energies are the full energies with unscaled Lennard-Jones radii. This increases the conformational freedom and therefore sampling for the short, at a cost of decreasing the acceptance probability in the outer loop. Scaled Lennard-Jones radii were also implemented in multiple time dynamics, but yielded very little apparent improvement because of the lack of phase space overlap between dynamics with different scaled Lennard-Jones radii). Rigid body moves are used to sample the translational and rotational space of the ligand. Multiple attempts with reduced VDW radii are applied, as it is quite common to fail in searching for a clash-free conformation in a single rigid body moving attempt (especially when the ligand is large and flexible and the binding pocket is relatively small). Each rigid body move includes 1000 attempts, and each attempt performs a translation along a random vector and a rotation around a random axis, with less than 0.5 angstroms and 60 degree displacement, respectively. In addition, the VDW radii are reduced (scaling factor 0.8) to soften the Lennard-Jones potential, so that mild steric clashes are allowed, which are likely to be resolved by the subsequent minimization. The rigid body move usually takes 20 to 40 seconds per move.

For rigid body Monte Carlo, a steric screen with an overlap factor of 0.7 was used, with a translation size of 0.5 Angstroms and a rotation size of plus or minus 60 degrees. No flip moves were included, as flips were not anticipated with the geometry of the ligand system [Robert, check this is true?] A Lennard-Jones scaling parameter of 0.8 was used during the inner steps. 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.

3. Hybrid Monte Carlo Moves: Hybrid Monte Carlo (HMC) is built on top of a molecular dynamics (MD) integrator. We implemented velocity verlet MD using the RESPA formalism [Tuckerman *et al.*, 1991], with bonded and short range interactions evaluated every inner time step, with long range interactions evaluated every long time step. We also included in the including Andersen, Berendsen, and Langevin thermostats, as well as Brownian Dynamics. Short range interactions included all bonded interactions and nonbonded interactions less than a user-specified cutoff. Both short and long range cutoffs are dipole-based as described in a previous paper [cite]. Energies are conserved in vacuum simulations if the long range cutoff extends to cover the entire molecule. With implicit solvent dynamics, energies are not conserved as there is no derivative of the Born alpha with respect to position, so a thermostat must be used to approximate realistic dynamics. One mode of action is to run HMC on top of normal RESPA dynamics, with acceptance and rejections done after a fixed number of steps of MD, as in typical HMC. Velocities are re-randomized after every Monte Carlo step from the Maxwell-Boltzmann distribution. However, HMC can also be implemented in a multiple-time step Monte Carlo framework [Hetenyi *et al.*, 2002]. In this case, molecular dynamics is run only using only the short range energies, which will conserve energy and thus yield proper Markov chain behavior while the cutoffs and Born alpha are not updated. Inner loop Monte Carlo steps are performed with only the short range dynamics, and outer loop Monte Carlo steps are performed with all interactions. In the case of implicit solvent simulations, the Born alpha are only reevaluated in the outer loop Monte Carlo steps. This allows for evaluation of the full surface integral relatively infrequently, and yields a properly weighted Boltzmann

distribution of conformations without needing position derivatives of the Born alphas. The hybrid Monte Carlo (HMC)[Duane *et al.*, 1987] move in PLOP performs simultaneous sampling for the selected residues in the protein side chains and backbone as well as the ligand. Each HMC move performs a 5 picosecond, constant energy molecular dynamic (MD) simulation (starting at 900K) on all the atoms in the selected residues. Taking up to 15 minutes per move, the HMC is the most expensive among all three types of moves in PLOP.

The underlying molecular dynamics for the hybrid Monte Carlo consisted of 500 5 ns RESPA [Tuckerman *et al.*, 1991] steps, each with 5 1 ns short range steps. The dynamics were initialized from a Maxwell-Boltzmann distribution at 900 K, and then run without any thermostat the full 2. 5 ps run. Minimization was performed after the single HMC step, with acceptance/rejection performed at a Monte Carlo temperature of 900 K.

2.2.3 Evaluation

2.3 Results

2.4 Discussion

Chapter 3

Other Improvements in the Protein Local Optimization Program

3.1 Regression Testing

A number of common problems inherent in large programming projects are identifying bugs or regressions One of the largest problems [Evans, 2013; Atwood, 2009] Cathedral Bazaar [Raymond, 1999]

3.2 Small Molecule Library

3.3 Crystal Symmetry

3.4 Knowledge Based Backbone Dihedral Penalty

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.

[Atwood, 2009] Jeff Atwood. Paying down your technical debt. `http://www.codinghorror.com/blog/2009/02/paying-down-your-technical-debt.html`, 2009.

[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.

[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.

[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.

[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.

[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.

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

[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.

[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.

[Evans, 2013] Jon Evans. Technical debt will kill you dead (if you let it). `http://techcrunch.com/2013/02/09/technical-debt-will-kill-you/`, 2013.

[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 multipole 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.

[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.

[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.

[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.

[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.

[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.

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

[Hetenyi *et al.*, 2002] Balazs Hetenyi, Katarzyna Bernacki, and BJ Berne. Multiple time step monte carlo. *The Journal of chemical physics*, 117:8203, 2002.

[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.

[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.*, 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.

[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.

[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.

[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.*, 2011] 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.

[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*.

[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.

[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.

[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 approximation. 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.

[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.

[Raymond, 1999] Eric Raymond. The cathedral and the bazaar. *Knowledge, Technology & Policy*, 12(3):23–49, 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.

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

[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.

[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.

[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.

[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.

[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.

[Tuckerman *et al.*, 1991] Mark E Tuckerman, Bruce J Berne, and Glenn J Martyna. Molecular dynamics algorithm for multiple time scales: Systems with long range forces. *The Journal of chemical physics*, 94:6811, 1991.

[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.

[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.

[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.

[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.