Biophysical Properties of Small Molecules Binding to Proteins

by

Richard D. Smith

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biophysics)
in The University of Michigan
2010

Doctoral Committee:

Associate Professor Heather A. Carlson, Chair
Professor Gordon M. Crippen
Professor E. Neil G. Marsh
Professor Shaomeng Wang
Assistant Professor Oleg V. Tsodikov

ACKNOWLEDGEMENTS

I would like to thank my advisor:

Heather A. Carlson, Ph.D.

I would to thank all the people who have helped create and maintain Binding MOAD:

Mark Benson, Ph.D., Nickolay Khazanov, James B. Dunbar Jr., Ph.D., Michael Lerner, Ph.D., Liegi Hu, Ph.D., Jason Nerothin, M.S., and Jayson Falkner, Ph.D.

Torrey Path:

Peter Dressler, Brandon Dimcheff, and John Beaver

I would like to thank my sources of funding:

Deans Fellowship, and The Molecular Biophysics Training Grant, NIH and NSF

I would like to thank all the members of the Carlson Lab, past and present:

Katrin Lexa, Jerome Quintero, Peter Ung, Kristin Meagher, Ph.D., Kelly Damm, Ph.D., Anna Bowman, Ph.D., Joslyn Y. Kravitz, Ph.D., Xiao-Jian Tan, Ph.D., Jeff Wereszczynski, Ph.D., Steven A. Spronk, Ph.D., Haizhen Zhong, Ph.D., and M. N. Jagadeesh, Ph.D.

I would also like to thank my committee:

Gordon Crippen, Ph.D., Neil Marsh, Ph.D, Oleg Tsodikov, Ph.D., Shaomeng Wang, Ph.D., Ioan Andricioaei, Ph.D. (former member), and Jignesh Patel, Ph.D. (former member)

I would like to thank Jignesh Patel, Pharm.D., and Bruce Mueller, Pharm.D. for their help with SAS.

I would like to thank Allen Bailey for keeping the computers up and running.

I would like to thank my family and friends.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

LIST OF APPENDICES

Appendix

ABSTRACT

Binding MOAD (Mother of All Databases) is the largest collection of high-quality, protein-ligand complexes. Binding MOAD contains 13,138 protein-ligand complexes comprised of 4078 unique protein families and 6210 unique ligands. We have compiled binding data for 4146 of the protein-ligand complexes. The creation of this database and three studies mining the database for biophysical properties of protein-small molecule binding are discussed in this thesis. An additional study is included in the appendix which investigates flexibility upon small molecule binding to MDM2.

First, we present the development of GoCav, which allows us to mine properties of the whole database. We have determined that most complexes have well buried binding sites (70-85%), which fits the idea that a large degree of contact between the ligand and protein is significant in molecular recognition.

Secondly, we investigate the differences in biophysical properties of binding to enzymes versus non-enzymes. Differences in the sizes of weak versus tight ligands indicate that the addition of complementary functional groups may improve the affinity of an enzyme inhibitor, but the process may not be as fruitful for ligands of non-enzymes. Non-enzymes were found to have greater ligand efficiencies than enzymes, which supports the feasibility of non-enzymes as druggable targets. This has significant ramifications for target selection in drug design. Most importantly, the differences in ligand efficiencies appear to come from the pockets which yield different amino acid compositions, despite similar overall distributions of amino acids.

We then investigate the biophysical properties of the most efficient protein-ligand complexes. All highly efficient small molecules contain one or more charge and are

found in binding sites with at least one charge, challenging previous thoughts that hydrophobic properties of ligands lead to the better binding. Lastly, it is known that affinity for complexes rarely exceeds -15 kcal/mol, and we suggest that ligands do not exceed this value because there is no evolutionary pressure to drive tighter binding.

CHAPTER I

Introduction

Proteins utilize the binding of small molecules to perform a wide range of biological functions. Common functions include protein processing, cell signaling, responding to environmental conditions, regulating and performing metabolism. Due to their great diversity of structure and function, proteins bind a vast array of small-molecule ligands. The precise biochemical and physical properties which significantly impact protein-ligand binding are of some debate. However, it is important to understand the contribution of these properties because structure based drug design relies on this information to develop small molecules which are able to bind to a particular target and create a desirable physiological response.

This dissertation utilizes a large database of high-quality, x-ray crystal structures of protein-ligand complexes, annotated with binding data, to determine what properties are important for small molecule binding. The creation of the database is discussed first. Three studies regarding the mining of the database are then presented. The first study details characteristics of protein-ligand binding as a whole in the database. The second provides insights into particular aspects of binding that are important to particular classes and families of proteins. The third study shows the biophysical characteristics of the complexes that exhibit the "most efficient" binding. These studies are based on a large database of static crystal structures, so to round

out these studies, the appendix presents molecular dynamics simulations to examine ligand binding to the MDM2 protein.

1.1 Protein Ligand Binding

Over time, theories have changed on how small molecules interact with a binding site of a protein. In 1894, a "lock and key" model was proposed by Herman Emil Fisher, where a protein pocket is preformed to fit a particular shape (3; 4; 5). A similar concept was suggested by Linus Pauling who indicated that active sites were preformed to fit the transition state of a reaction, rather than the substrate or ligand (6; 7). This theory was later updated by Daniel Koshland. He introduced an induced-fit model, where a protein would change to adapt to bind a small molecule (8). However, more recent evidence suggests that proteins exists in an ensemble of structures, including ones that resemble the bound state of the protein, and upon binding of the small molecule, the population distribution will shift to favor the bound state (9; 10; 11; 12; 13).

Of utmost importance to researchers is how tightly a small molecule binds to a protein. The free energy of binding is defined by, entropy (ΔH) and enthalpy (ΔS), in the following relationship: $\Delta G_{binding} = \Delta H_{binding}$ - $T\Delta S_{binding}$ = -RTln(K_A), where K_A is the equilibrium constant of the binding between protein and ligand. The precise contribution of enthalpy and entropy into the calculation of the free energy of binding (ΔG) is dependent on the protein and small molecule under investigation. Several factors are involved in determining both the enthalpy and the entropy involved in protein-ligand binding (14; 4), but these values are inherently difficult to calculate accurately. First, it is based on determining a small difference between two very large numbers, the energies of the complex and the energies of the protein and ligand alone interacting with solvent. Secondly, the entropic contributions are difficult to estimate because the conformational space available to both protein and small

molecule is potentially large. Lastly, many different factors must be considered: van der Waals interactions, electrostatic interactions, (de)solvation, and flexibility of both the protein and small molecule ligand (14). Here, we aim to circumvent these limitations by using our large database of protein crystal structures solved with bound small molecules to determine what types of contacts lead to the best binding. We will also look at the ability of different classes and families of proteins to investigate any possible differences in their ability to bind small molecules.

1.1.1 van der Waals Interactions and Electrostatic Interactions

One of the most significant contributions to the binding affinity has been thought to be van der Waals interactions (15). These are low-energy interactions created by the London Dispersion forces arising from placing atoms in contact with each other. The shape complementarity of the binding pocket to the small molecule allows for optimal contacts, in agreement with the "lock and key" model (16). Most small molecules bound to proteins are well buried to maximize the amount of contact being made. Liang *et al.* found that binding sites are buried cavities or have one or two small exposed areas and that binding pockets tend to be the largest pockets in the protein (17).

On the other hand, electrostatic interactions have much stronger enthalpic contributions and include hydrogen bonds, contacts to metals, and salt bridges. Hydrogen bonds form between highly electronegative atoms (generally O or N) and a hydrogen bound to another highly electronegative atom. These bonds generally contribute 3-7 kcal/mol to the enthalpy (18). However, the precise contribution depends upon the geometry of the hydrogen bond (4). Hydrogen bonding interactions are also made with water, leading to a large desolvation penalty for both the ligand and binding site; therefore, it is believed they generally do not contribute much to the free energy of binding, since it is the difference between the standard free energy unbound in

water and the standard free energy of the complex (4). For example, Lafont *et al.* tried to gain a greater free energy of binding by adding a functional group to an HIV protease inhibitor to make an extra hydrogen bond between the ligand and the protein. Although a gain in enthalpy was achieved, it was completely compensated by entropic loss induced by both desolvation of the polar group and forcing that polar group into a particular conformation (19). Salt bridges are made between positively and negatively charged functional groups, and are the strongest non-covalent interactions that can be made. However, the desolvation penalty of removing water from a charged group is also quite large (20).

There has been contradictory evidence as to which types of interactions play the most significant roles even in the strongest known natural protein-ligand complexes, specifically in the binding of biotin to streptavidin, the tightest known natural complex. In 1993, Miyamoto and Kollman used free energy perturbation calculations on biotin-streptavidin and N-L-acetyltryptophanamide-α-chymotrypsin to show that the increased binding affinity for the biotin-streptavidin system can be accounted for by van der Waals contacts made in the biotin-streptavidin complex where the pocket in streptavidin is preformed as in the traditional lock and key theory (21). However, newer work using combined quantum mechanics/molecular mechanics and monte carlo computational techniques on hydrogen-bonding residues in streptavidin have indicated that networks of hydrogen bonds are responsible for the strong binding in the biotin-streptavidin complex (22). The importance of the network of hydrogen bonds was also confirmed using isothermal calorimetry, which showed an 11-fold greater contribution to the free energy of binding of two coupled residues involved in hydrogen bonding than the contribution of each of the two residues individually (23).

A common metric to evaluate how well a small molecule binds is "ligand efficiency". This metric is defined as the binding affinity per number of non-hydrogen atoms (24; 25; 26). It was first introduced by Kuntz *et al.* in 1999 (27). Kuntz ana-

lyzed 159 complexes and scaled the affinities by the number of non-hydrogen atoms present in a ligand as a metric of size-independent affinity. They showed that for each non-hydrogen atom the most to be gained is ˜-1.5 kcal/mol of binding affinity (27). This maximum was consistent with their theoretical predictions based on van der Waals and hydrophobic interactions (27). The hydrophobic effect will be discussed further below.

Although electrostatics have previously been proposed to have little effect on the free energy of binding, Bruce Tidor has been working to optimize charge complementarity to improve binding free energy (28; 29; 30). He has developed an analytical solution to the Poisson equation to model the electrostatics of a binding site, and an analytical method of optimizing the charge profile of a ligand, taking into account the desolvation penalty, to match the calculated electrostatics of the binding site (28; 29; 30). His group was able to predict a position to improve the charge complementarity of a small molecule bound to chorismate mutase. They suggested that this improvement would lead to a 2-3 kcal/mol benefit to the free energy of binding (31). However, this was not confirmed experimentally. The effects of hydrogen bonds and electrostatics have been shown to be dependent on distance, short-range hydrogen bonds of less than 2.5 $\mathring{A}$ have been shown to lead to binding affinities of greater than 15 kcal/mol (32; 33; 34). This contradicts the previous idea of a diminished role of electrostatics in small molecule binding. In chapter five of this thesis, we present results based on distances of charge-charge interactions and their impact on the efficiency with which small molecules bind.

1.1.2 Desolvation and Solvation

The binding of small molecules to proteins occurs in an aqueous environment, so water plays a significant role in the binding. Water must be removed from the binding site as well as the ligand (4). If water molecules are found in the binding

site, they are generally partially occupied and are able to move in and out of the pocket. Desolvation can be either favorable or detrimental to small-molecule binding. Desolvating a charged functional group is unfavorable to binding (35), whereas the "hydrophobic effect" results in desolvation being favorable to binding.

The "hydrophobic effect" was first proposed in 1945 by Frank and Evans and is a positive influence on the free energy of binding (36). The placement of a non-polar molecule in water is an energetically unfavorable process (37). The hydrogen-bonding network of water becomes disrupted locally around the non-polar molecule. Therefore, burial of non-polar groups within the protein is seen to have a positive effect as water is able to rearrange back to its favorable interactions with itself (38). It has been shown that reorganization of the solvent can attribute anywhere from 25 to 100% of the enthalpy gained in small-molecule binding (39). It has also been shown that the enthalpic contribution of the hydrophobic effect is proportional to the amount of buried non-polar surface area (38).

1.1.3 Ligand and Protein Flexibility

The formation of van der Waals interactions and the "hydrophobic effect" have been seen to have favorable impact on the free energy of binding, while the impact of electrostatics is dependent on the precise protein-ligand complex. There are other factors working against protein-ligand binding. The loss of flexibility of the small molecule and the protein by forcing them into a particular conformation is a loss of entropy and thus a penalty in the free energy of binding. The loss of rotation in side chains upon small molecule binding has been estimated to be ˜0.88 kcal/mol per residue (40; 4; 41). It is also important to note, that in some cases some protein residues will increase in flexibility, such is the case in Topoisomerase 1 (42). Also, NMR studies have shown that backbone flexibility can increase upon ligand binding to the mouse urinary protein (43). Additionally, Yang *et al.* found that in a set

of 63 complexes from the Protein Data Bank 29 % of the atoms in the binding site became more flexible than the corresponding free structure, using the B-factors as a metric of flexibility (44). While, not all structures had atoms that displayed increased flexibility, 75 percent of the structures had at least some portion of the molecule that had some increase in their B-factors (44). To ensure the B-factors are comparable between systems, the structures used in the study were solved by the same group and had resolutions less than 2.5 $\mathring{A}$ and R-values less than 0.245, with the exception of two structures (44). A recent study by Yang *et al.* has noted the importance of small molecule reorganization in the prediction of binding affinity using a wide range of scoring functions (45). Different orientations of the small molecule binding in the binding site is observed in the case of camphene, adamantine, and thiocamphor bound to cytochrome $P450_{cam}$ (46).

1.2 Surface Area Calculations

In this thesis we utilize GoCav to calculate the surface area. GoCAv was developed along with this thesis work in order to calculate the buried surface area of a small molecule based on its location in an x-ray crystal structure. It also can handle binding sites that are exposed, an area in which previous methods have been unable to calculate surface areas accurately. Other programs that have been developed for this are POCKET(47), SURFNET(48), CAST(17), PASS (49), and NACCESS(50).

1.2.1 POCKET

POCKET was developed by Levitt and Banaszak in 1992. This program identifies pockets by scanning a grid on the x, y, and z-axis to find where a probe of a certain radius does not touch any protein atoms. The surfaces are then determined using a variant of the marching cube algorithm, in which a surface cube is determined by the surrounding cubes. The shape of the surface is determined by a set of triangles

associated with the cubes (47). This program will find all pockets in the protein regardless of size, but does not specifically provide the specific pocket of the desired ligand, although it may be one of the pockets found.

1.2.2 SURFNET

SURFNET was developed by Laskowski in 1995. SURFNET generates the surface by adding a Gaussian density function about the center of each atom. At a specific contour level, the atomic spheres are generated and spheres are placed in between atoms to find gap regions. The binding pockets are then considered the largest of the gap areas. This also may not find highly exposed binding pockets, since it needs two protein surfaces to determine the gap areas (48).

1.2.3 CAST

CAST was developed by Liang et al. in 1998. It uses Veronoi tessellations to map out the surface of the binding site. The tessellations are formed from triangles (tetrahedral in 3-dimensions) created by using the atoms of the protein as vertices. Triangles that do not contain any other atoms are considered cavities. This program also cannot find exposed binding sites as the number of triangles will go to infinity for exposed binding pockets (17).

1.2.4 PASS

PASS was developed in 2000 by Brady *et al.* In this algorithm probes are placed on the surface of the protein using triplets of protein atoms, then searches for points where a probe sphere can lie tangential to all three atoms. Any probes that are exposed to the solvent are then removed. This process is continued by placing another layer of spheres on the surface of the spheres placed in the first iteration. The algorithm ends when no probes can be removed. Clusters of four or more probes are

kept. PASS does not calculate the surface area and will not identify exposed binding sites, since if the probes are exposed to the surface they are automatically removed and no more probes spheres are placed in that area (49).

1.2.5 NACCESS

NACCESS was developed by Hubbard and Thornton in 1993 (50). It calculates the solvent accessible surface area based on Lee and Richard's method developed in 1971 (51). This method uses intersections of atomic spheres with their van der Waals radii to create the surface of the protein. Planes are drawn through the intersections and the remaining convex arcs are obtained as the van der Waals surface. The solvent accessible area is created by augmenting the atomic radii with a probe radius. This program does not locate a binding site and merely calculates the surface area of any atom included in the calculation (50).

1.3 Protein-Ligand Databases

Given the discussion regarding the thermodynamics of protein-ligand binding, it is important to investigate a wide range of proteins bound to a variety of small molecules. It is not surprising that different proteins will have different contributions to the free energy of binding. A large database of these interactions is necessary because we aim to make generalized statements regarding a wide range of diverse complexes. The best source of protein structures is available from the Protein Data Bank, where greater than 60000 structures are deposited (52; 53). It is also important to have structural coordinates of complexes correlated to the experimental binding affinity. Other databases of protein-ligand interactions with binding data have been created prior to this work, namely Ligand-Protein Database (LPDB)(54), BindingDB(55) (although it does not contain coordinates it has a large number of binding affinities reported with links to some protein crystal structures in the PDB), Protein Ligand

Database (PLD), and PDB Bind(56; 57). However, all have their deficiencies with respect to combining binding affinity and atomic coordinates. The next section discusses these databases in more detail. Other databases, such as MSDsite(58) and Relibase+(59; 60) only have protein-ligand complexes, and do not have affinities.

1.3.1 LPDB

The Ligand-Protein Database (LPDB), created in 2001, has 195 complexes with binding data. LPDB also provides computer generated docking decoys to help researchers in developing more accurate scoring functions. LPDB has been analyzed to address redundancy of the protein structures. The 195 complexes consist of 51 unique proteins in 21 protein classes (54). LPDB was created using complexes found in training sets of previously used scoring functions and searching for those complexes in the Protein Data Bank (54).

1.3.2 Binding DB

Binding Database (Binding DB), also created in 2001, contains very high-quality thermodynamic data for 722 proteins. Binding DB also accepts the deposition of Ki data, and the number of entries has grown significantly to 62,134 binding reactions (http://www.bindingdb.org/bind/stat.jsp) and continues to grow. Most of the data is now inhibition constants. Binding DB's strength lies in the volumes of information given on experimental conditions used in determining binding information, including raw data in some cases. Most of the ligands do not have a pdb structure, but at least one structure of the protein bound to some ligand exists (55).

1.3.2.1 PLD

The Protein Ligand Database (PLD) by John Mitchell in 2003 is a small database of protein-ligand complexes (61). All of the entries are annotated with calculated

binding energy using the knowledge-based method BLEEP. 357 entries are annotated with experimental binding data. While ligand similarity scores have been calculated, they are not available (61).

1.3.3 PDBBind

PDBbind, created in 2004, contains binding data on 3214 complexes, with 2084 unique ligands, collected from the PDB (56; 57). PDBBind was curated in a very similar fashion as our database, but has some key differences. PDBBind focuses on complexes with only one ligand in the crystal structure (56; 57). PDBbind also excludes any complex binding a simple cofactor such as ATP. Our database does not discriminate against molecules such as ATP, since it is also a small molecule that binds to a defined binding pocket in proteins. PDBBind has no threshold value for quality of the electron density(the largest crystal structure resolution is 4.7 $\mathring{A}$). PDBbind only provides structures of complexes for which it has binding data (56; 57). To meet our specific needs, we have created our own database, Binding MOAD (Mother of All Databases) (62).

1.4 MDM2

The majority of the thesis will discuss the creation of Binding MOAD to answer the questions as to what biochemical properties lead to high-affinity binding. However, we do not take in to account system flexibility. Therefore, we present a study regarding the highly flexible MDM2 protein and investigate the flexibility of the protein using molecular dynamics upon binding of a small molecule. The next session discusses the importance of MDM2 and the details of the flexible binding site.

The p53 tumor suppressor, also known as the guardian of the genome, is vital in cell cycle regulation, DNA repair, and apoptosis (63; 64; 65). Mutations in p53 are seen in approximately half of all human cancers (66). Where p53 is in wild-type

form, it is inhibited by over-expression(67; 68) or amplification(69) of murine double minute 2 oncoprotein (MDM2; also referred to as HDM2 in human). Reactivation of p53 through inhibition of the p53-MDM2 interaction has been shown to be a novel approach for initiating or enhancing cancer cell death (70; 71). A better understanding of MDM2 dynamics is important for the design of more selective and potent inhibitors of the MDM2-p53 interaction.

A crystal structure containing residues 25 to 109 of MDM2 and residues 17 to 29 of p53, was solved in 1999 (1YCR) (72). This showed two approximately similar subdomains, which come together to form a binding cleft for p53. Three side-chains of p53 (Phe19, Trp23, and Leu26) fill the relatively deep hydrophobic pocket of MDM2. This crystal structure has been the basis of several dynamics studies (73; 74; 75; 76), in all cases the authors compared the MDM2-p53 complex to *apo*-MDM2, which was generated by removing the peptide.

Barrett *et al.* utilized *CONCOORD*(77), a non-Newtonian method of ensemble generation to examine protein motion in creation of their program Dynamite (73). They found that the principle mode of *apo*-MDM2 was a bilobal flexing, or breathing, of the protein; this motion was greatly reduced in the p53 bound complex. Previous work in the Carlson lab has utilized MD simulations to develop receptor-based pharmacophore models. The models were used to identify five small-molecule inhibitors of the MDM2-p53 interaction (78; 76). Espinoza-Fonseca and Trujillo-Ferrara presented two 35-ns molecular dynamics (MD) simulations; again demonstrating that the *apo*-MDM2 had a highly flexible and narrow cleft (75). Whereas with p53 bound, the cleft was more stable and wider. They also reported important side-chain motions in residues Leu57, Tyr67, His96, and Tyr100 which were present in *apo* MDM2 but not MDM2-p53, and they suggested that these motions are involved in the molecular recognition of p53 and other ligands (75).

The deep, well-defined binding cleft shown from in the crystal structure of MDM2-

p53, suggested that the MDM2 cleft would be a suitable target for small molecule inhibitors. To date, several small molecule inhibitors of the MDM2-p53 interaction have been reported (reviewed in (79; 80; 81)). The crystal structure of MDM2 was solved with both a member of the nutlin class (1RV1)(82) and from the 1,4,-benzodiazepine-2,5-diones (1T4E) (83). Several other structures have been solved with a variety of small molecules, in all, there are ten structures solved bound to a ligand.

The sequence of MDM2 residues 16-24 is highly conserved in mammals (84). NMR studies show that these residues form a "lid" which stabilizes MDM2 in the absence of p53 (84; 85). When the lid is closed, it shields the hydrophobic binding cleft of MDM2. Ile19 occupies the same space as Pro27 of bound p53, and makes interactions with His96, Arg97, and Tyr100 (85). However, the lid is easily displaced 3-4 $\mathring{A}$ to deepen the binding cleft and then peptide or inhibitor completely binds (84; 86; 85). In the appendix, we present work in progress to examine the role of the lid and the flexibility of the system during ligand binding.

1.5 Conclusion

The precise biophysical characteristics that determine the affinity with which a small molecule binds to a protein is highly variable. Many believe that the primary interactions favorable to binding are van der Waals contacts and desolvation due to the "hydrophobic effect". However, electrostatics-such as hydrogen bonding and charge complementarity-have been shown to also have a favorable impact, despite the higher desolvation penalty and the fact that these interactions are also made with water. Although flexibility and conformational entropy play roles in the free energy of binding, it is very difficult to account for in calculations. Since we are using a large database of protein-small molecule complexes from static x-ray crystal structures, this thesis is rounded out by a molecular dynamics investigation of the highly flexible protein MDM2.

In the first chapter, we discuss the creation and curation of Binding MOAD (Mother Of All Databases). The second chapter discusses the development of GoCav for calculating surface areas and provides some general trends regarding the complexes in Binding MOAD. The third chapter breaks down the database into families and notes differences in the physical properties of the protein-ligand complexes, namely enzymes versus non-enzymes, as well as tightly bound ligands versus weakly bound ligands. The fourth chapter investigates the complexes that have the most efficient binding, based on an affinity per atom or per buried surface area metric. The appendix provides preliminary results regarding the binding of a specific small molecule to the human MDM2 protein, to investigate the role of flexibility on a protein-small molecule complex.

This thesis work has far reaching implications for computational biology and theoretical biophysics. In several reviews of methods of structure-based drug design, each points out the need for databases which provide structural data of protein-ligand complexes as well as binding affinity in order to provide training sets and tests sets for scoring functions. These reviews cite Binding MOAD and databases like it as valuable resources for improving docking and scoring algorithms (87; 88; 89; 90; 91; 92). Since publishing Binding MOAD, many researchers have acknowledged the usefulness of Binding MOAD, but have created similar databases that have additional information regarding the ligand and/or binding site, and/or have provided binding affinities for complexes which do not have structures deposited (93; 94; 95; 96; 97; 98). These databases are publically available: AffinDB(93), sc-PDB(94), SuperSite(95), PDBCal(96), PLID(97), and PSMDB(98). Others have also acknowledged Binding MOAD, yet have created datasets to meet their specific research aims(99; 100; 101). Potential uses of Binding MOAD have been suggested to be of benefit in two projects. First, it can be used to augment the DUD dataset, which is a dataset of decoy ligands (102), and as a link from the cheminformatic toolkit developed by Rosania *et*

al. (103). Additionally, Binding MOAD has been noted to be unique in its inclusion of ligands bound to heme containing proteins (104).

Binding MOAD has also had some implications in research that is ongoing in investigations of binding sites. Park and Kim utilized the ideas of "invalid" ligands, developed in Binding MOAD to create a dataset of ligand binding sites from the PDB. They utilized this dataset to link structure to function by creating a network model of similar binding sites (105). Daily and Gray also used this idea to create a dataset in their investigation of conformational changes in allosteric proteins (106).

Binding MOAD was directly used in four studies. A subset of Binding MOAD has been used to investigate the specificity of binding of FAD and NAD (107). Binding MOAD was used in conjunction with AffinDB, PDBBind, and PLD to develop a training set to evaluate a model for predicting the affinity of enzyme-ligand interactions(108). Binding MOAD was also utilized to locate the binding affinites of specific molecules used in their study of Fluorine containing compounds bound to proteins (109), and yet another study used it to find a small chelating compound that binds to an antibody with high-affinity (110). Lastly, the development of Binding MOAD has helped lead to the formation of the Community Structure Activity Resource (CSAR) center, which is the only NIH funded center designed to gather data to improve scoring of protein ligand complexes for structure based drug design.

The investigations of binding sites as a whole, as well as enzymes versus non-enzymes is beginning to influence how we look at binding sites. In a recent review of drug discovery for protein-protein interactions, our research has provided a limit to the size of protein-ligand binding sites compared to protein-protein interfaces (111). The study of enzyme binding sites versus non-enzyme binding sites has helped to shape our knowledge of enzyme binding sites (112). It has also provided data that indicates the necessity for different strategies for improving binding to enzymes and non-enzymes, as well as data that point to non-enzymes being more "druggable".

This has potential to effect the choice of protein targets in the drug industry, since they may be inclined to use a more druggable target to increase their chance of success. Our investigation is beginning to change how the community looks at ligand efficiencies and suggests that it cannot be applied strictly when investigating multiple systems (113). This has potential implications in how we think about designing small molecules for different types of proteins.

The study of the most efficient ligands implies that short electrostatic interactions of very small molecules are responsible for the most efficient small molecules. This may indicate the desolvation penalty of highly charged molecules is not as large as previously thought. We also suggest that affinities better than -15 kcal/mol may be difficult to attain because there is no evolutionary driving force to allow this selection. When ligands have such high affinities, their bound lifetimes are on the order of days to weeks. Many proteins degrade before these ligand are able to dissociate, and may explain the limit of -15 - -19 kcal/mol. The fact that short electrostatic interactions define the boundary of ligand efficiency challenges previous ideas of the biophysics of small molecule binding, where hydrophobic and van der Waals are the dominant interactions which lead to improved binding.

CHAPTER II

Binding MOAD (Mother of All Databases)

2.1 Introduction

Binding datasets for protein-ligand complexes were first used in computational chemistry to develop scoring functions for ligand docking and de novo design of enzyme inhibitors. The earliest relevant dataset was only 45 complexes(114) and more recent sets are 200-800.(54; 55; 56) Some sets have been made available online, changing their nature from a flat list of data in a paper to a dynamic and searchable tool for the scientific community. The largest and most useful datasets are outlined below. The strengths of each are noted and the comparative strengths of Binding MOAD are highlighted. Our aim is to make Binding MOAD the largest possible collection of high-quality, protein-ligand complexes available from the Protein Data Bank (PDB)(115) and augment that set with the inclusion of binding data. When initially introduced in 2005, Binding MOAD contained 5331 protein-ligand complexes, of which binding data was collected for 1375 (26%) of the protein-ligand complexes. As the PDB grew, we have updated the dataset three times. Currently BindingMOAD contains 13,138 structures, with binding data available for 4203 (32%) of these structures. The numbers presented in the following text represent the current state of Binding MOAD.

2.1.1 LPDB

The Ligand-Protein Database (LPDB) has 195 complexes with binding data.(54) LPDB also provides computer generated docking decoys to help researchers in developing more accurate scoring functions. We do not plan to add decoys to Binding MOAD, but our dataset is an order of magnitude larger. LPDB has been analyzed to address redundancy of the protein structures. The 195 complexes consist of 51 unique proteins in 21 protein classes.(54)

2.1.2 Binding DB

In one of the first papers announcing the Binding Database (Binding DB), it was reported to contain very high-quality thermodynamic data for 400 binding reactions (90 for biopolymers).(55) Binding DB has recently started to accept the deposition of K_i data, and the number of entries has grown significantly to >60,000 binding reactions (http://www.bindingdb.org/bind/stat.jsp). Most of the data is now inhibition constants for biopolymer binding. Binding DB's strength lies in the volumes of information given on experimental conditions used in determining binding information, including raw data in some cases. Though we do not provide isothermal titration calorimetry details like Binding DB, our dataset is larger and we supply structural data from the PDB. The complexes in Binding DB are not cross-linked to their structural data.

2.1.3 PDBbind

PDBbind was created by Shaomeng Wang and coworkers.(56) It contains binding data on 2665 complexes with resolution 2.5 Å (459 structures > 2.5 Å are also provided as a secondary set). PDBbind does not address redundancy, but does note that approximately 200 different types of proteins are present. This set was curated in a similar fashion as Binding MOAD but focuses on complexes with only one ligand

in a pocket. PDBbind also excludes any complex binding a simple cofactor such as ATP. Binding MOAD is larger because we do not ignore cofactors or protein-cofactor-ligand complexes. We also provide information on the structures when we do not have binding data because they are still a valuable resource in database mining. PDBbind only provides structures of complexes for which it has binding data.

PDBbind and Binding MOAD were developed independently at the University of Michigan, Ann Arbor. When we learned of our similar research efforts, we found that our goals were synergistic. The research projects around PDBbind focus on developing scoring functions and searching ligand substructures. Our focus with Binding MOAD is more on protein binding sites and protein flexibility. In sharing binding data between our groups, we found a disagreement of only 1%, which highlights the high accuracy and quality of binding data collected in both groups. Disagreements were simple typos that were easily corrected by consulting the reference again. This arrangement allows both groups to double check all of the data, basically eliminating the errors inherent in hand-processed data. This high level of quality control is unheard of for datasets of this size.

2.1.4 Other Online, Protein-Ligand Databases Without Binding Data

Of course, various improvements are constantly being added to the PDB to provide additional information and viewers to aid understanding protein-ligand complexes.(116; 117) However, several other online resources deserve discussion. These databases do not present binding data for the protein-ligand complexes in the PDB, but they do provide useful search tools, various analyses, and viewers of PDB complexes.

Relibase+ and MSDsite are similar datasets that specifically focus on protein-ligand complexes. In 2002, Relibase+ contained 15,454 PDB entries, 50,514 individual ligand sites, and 4530 unique ligands.(59; 60) MSDsite is the newest resource in the MSD suite of web-based tools from the European Bioinformatics Institute.(58)

However, the description of ligands in both datasets is unusual for our application. We have taken great care to make extensive lists of molecules to exclude as ligands in Binding MOAD. Metal cations like magnesium, inorganic salts such as sulfate, and common crystal additives like polyethylene glycol are not counted as ligands in Binding MOAD, but they are ligands in Relibase+ and MSDsite. They even count modified amino acids in the protein chain as ligands. The strengths of Relibase+ and MSDsite are that they provide powerful search tools for mining their datasets for interaction patterns. A benefit to the description of ligands in Relibase+ and MSDsite is that it allows a user to investigate a protein's interactions with a feature like a modified residue, a structural zinc ion, or an inorganic reactive center in the active site. These groups are simply considered to be part of the protein in Binding MOAD because of its focus on substrates, organic cofactors, and inhibitors. Such an investigation is not possible with Binding MOAD at this time.

PDBsum and MMDB do not focus on protein-ligand interactions, but they provide resources that are very useful for those interests. PDBsum is an online resource from Laskowski and Thornton(118; 119; 120) that provides analyses for all structures in the PDB (not just protein-ligand structures). PDBsum provides chemical, enzymatic, and genomic information about the entry, and it provides viewers to analyze protein-ligand interactions. The viewers display secondary structure, ligand interactions, and cavities. MMDB is Entrez's 3D-structure database.(121) Its focus is protein data, but several resources for comparing related sequence and structure have direct relevance for ligand binding.

2.1.5 Redundancy in Protein-Ligand Databases

Binding databases available to-date usually do not address the issue of redundancy. Many protein complexes have more than one bound structure. Many small datasets contain several examples of HIV protease, dihydrofolate reductase, thrombin,

trypsin, lysozyme, etc. To address this issue in Binding MOAD, we have analyzed for redundancy and grouped proteins by 90% sequence identity. Of 13,138 complexes in Binding MOAD, there are 4078 unique protein families when clustered at 90% identity. In our nonredundant version of Binding MOAD, each protein family is represented by the structure of the tightest binder. Of the 4078 complexes in the nonredundant set, we have obtained binding data for 1176. (In cases where binding data was not available, best resolution and other factors were used to choose representatives of the protein families). As we mine this database for general biophysical properties, our results for redundant and nonredundant Binding MOAD can be compared to measure the influence of bias in the structures available in the PDB. Also, inverse docking techniques, where a single ligand molecule is screened against a set of many proteins, will require a nonredundant set of protein complexes.(122; 123)

2.2 Methods

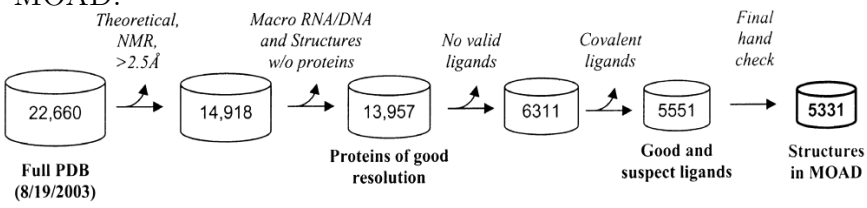
2.2.1 Top-Down Approach

Older protein-ligand databases were originally created by reading through the literature and compiling lists of appropriate complexes and their binding affinities. This sort of bottom up approach relies on finding good information in a relatively random fashion. We chose a top down approach to create Binding MOAD so that it contained every protein-ligand complex with a 3D structure. We started with the entire PDB,(115) removed inappropriate structures, and used the remaining structures to guide our literature searches in a systematic fashion. Since almost all protein structures are annotated with the authors' names and the appropriate reference, a starting point for the literature search is straightforward.

2.2.2 Paring Down the PDB

Perl scripts were written to determine whether each protein structure was an appropriate entry for Binding MOAD (Figure 2.1). Our scripts originally took advantage of the STAR parsers(124) from the Research Collaboratory for Structural Bioinformatics (RCSB) and the new mmCIF format from the uniformity project. The mmCIF files have gone through additional checks to correct sequence and EC errors that may exist in the legacy PDB files.(125) By using the mmCIF files, we plan to keep abreast of the newest improvements in data from the RCSB, making our resource more timely, accurate, and valuable. Since the uniformity project has not been continued, we now use the remediated PDB files, and have modified our scripts to parse these files using the Bioperl PDB parser. Our technique is similar to that used by Rognan and coworkers to create sc-PDB, a set of protein binding sites for inverse docking.(123) The major difference is that we did not use a keyword search to identify complexes. Our group and others have found that keyword searches miss complexes that can be identified through analyzing the individual structures. Starting with the entire PDB (22,660 structures on 8/19/2003), we eliminated theoretical models, NMR structures, and structures with poor resolution (> 2.5 Å). Large macromolecular complexes between proteins and nucleic acids were removed. However, we wanted to keep any metabolic enzymes that process nucleic acids, so structures with chains of four nucleic acids or less were kept in Binding MOAD. Short chains of 10 amino acids or less were counted as peptide ligands. Short-chain ligands were identified in the SEQRES section of the PDB format ({`_pdbx_poly_.seq_scheme`} data items in mmCIF format). Small molecule ligands were identified in the HET and FORMUL (in PDB format) sections ({`_chem_comp`} in mmCIF) or in ATOM and HETATM (in PDB format) ({`_atom_site`} in mmCIF). Initial filtering of the database utilized the mmCIF files from the uniformity project, however, currently we utilize the remediated PDB files.

Figure 2.1: Criteria to judge all PDB structures for entry into Binding MOAD. The scripts evaluate each structure - one at a time - against all criteria, but this step-by-step diagram is given to show the impact of each criterion. The numbers shown are taken from the first public release of Binding MOAD.



Covalently linked ligands were identified by calculating the minimum distance between the protein and each ligand. Minimum distances greater than 2.4 Å were defined as noncovalent. Values between 2.1-2.4 Å were examined visually to determine covalency. Distances less than 2.1 Å were considered covalent unless the short contact was to a metal ion (we considered many common catalytic metals to be part of the protein during this analysis). All short contacts to metals were examined visually. This was crucial in the case of zinc-containing enzymes where a zinc-ligand distance < 2.1 Å is not necessarily a covalent bond.(126) HET groups within 2 Å of another HET were identified as multipart ligands (unless they had partial occupancy and were actually two ligands occupying the same space). If any group of a multipart ligand was covalently linked to the protein, all components are identified as a covalent modification. This was important in the case of sugar chains on glycosylated proteins. Proteins with covalent modifications can still be part of the database if they have another acceptable ligand. If all ligands are covalent or inappropriate (see Table 2.1), the crystal structure is rejected.

2.2.3 Extensive Hand Curation of the Data

The literature citations for all final structures were read to confirm the validity of the ligands and find binding data. Our preference for affinity data is K_d over K_i over IC_{50}. Table 2.1 shows the great care that was taken to ensure that entries in

Table 2.1: Definition of Unusual HET Groups

Classification	Type of HET (Examples)
111 Suspect ligands	**Sugars** (glucose, galactose, fructose, xylose, sucrose, β-D-xylopyranose, trehalose) **Small organic molecules** (phenol, benzene, toluene, t-butyl alcohol) **Membrane components** (phosphatidylethanolamine, palmitic acid, decanoic acid) **Small metabolites** that may be buffer components (citric acid, succinate, tartaric acid)
78 Partial ligands	**Chemical groups** (amino group, ethyl group, butyl group, methoxy, methyl amine) **Inorganic centers** of transition state or product mimics (aluminum fluorides, beryllium fluorides, boronic acids) **Modifications to amino acids** (oxygens of oxidized CYS, phosphate group on TYR)
511 Rejected ligands	**Unknown or dummy groups** (UNK, DUM, unknown nucleic acid, fragment of) **Salts and buffers** (Na^+, K^+, CI^-, PO_4^{-3}, CHAPS, TRIS, tetramethyl ammonium ion) **Solvents** (DMSO, hexane, acetone, hydrogen peroxide) **Crystal additives and detergents** (polyethylene glycol, oxtoxynol-10, dodecyl sulfate, methyl paraben, 2,3 propanediol, pentaethylene glycol, cibacron blue) **Metal complexes** that associate to the protein surface and are used for phase resolution (terpyridine platinum, bis bipyridine imidazole osmium) **Metal ions** that are part of the protein (Mg^{+2}, Zn^{+2}, Mn^{+2}, Fe^{+2}, Fe^{+3}) **Catalytic centers** that are part of the protein (4Fe-4S cluster, Ni-Fe active center) **Heme groups** (heme D, bateriochlorophyll, cobatamin, protoporphyrin IX)

For brevity, not all compounds are listed.

Binding MOAD contain only appropriate protein-ligand structures. Short protein-ligand distances and suspect ligands were flagged for visual inspection in a more careful hand-check stage. Suspect ligands are crystal additives that are valid only in some cases. Partial ligands are molecules that cannot be a ligand on their own but are often a component of multipart ligands. Any HET with 3 heavy atoms is automatically part of this list. The covalency check identifies if these HET are modifications to the protein or a ligand.

The reason for our choice to reject or suspect various HETs in Table 2.1 is obvious in many cases. The reader may notice that β-D-N-acetylglucosamine (GlcNac, NAG in the PDB) is not on the suspect lists. We found that GlcNac was never used as a crystal additive. It was either part of a ligand or a covalent modification that was readily identified by our scripts.

Modifications to amino acids are on the partial ligand list because they can be part of the protein or part of a peptide ligand. Complexes containing heme groups were rejected because the covalent association of ligands to the central metals made it difficult for us to properly identify the true ligands. In many cases, it was a small molecule (oxygen, carbon dioxide). Of course, this neglects P450s which are very important in medicinal chemistry, toxicology, and pharmacology.(127) We plan to add P450s to Binding MOAD in the future to make it more useful.

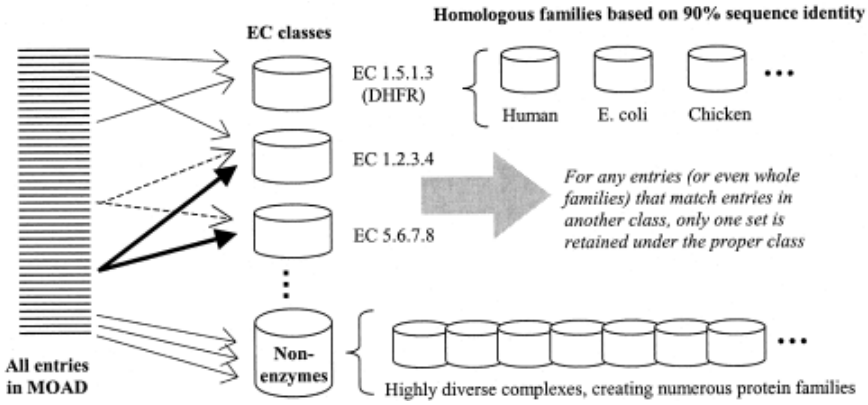
2.2.4 Grouping the Proteins to Address Redundancy in the Data

It is desirable to group proteins by related structure and function so that users can compare related systems. Enzyme classification (EC) numbers are used to broadly group entries into classes with similar chemical functionality. Within these classes, proteins are grouped into homologous protein families based on sequence.

The EC numbers and protein sequences are pulled from the mmCIF files of all appropriate structures. To compare the sequences in Binding MOAD, we use BLASTp

v2.2.7.(128) Defaults are used (E = 10, BLOSOM62 matrix, gap cost = 11, gap extend cost = 1). To create protein families, we use a cutoff of 90% sequence identity like HOMSTRAD,(129) but our grouping of proteins is slightly different than the clustering used for grouping similar sequences at the PDB.(130) The routine is presented in Figure 2.2:

Figure 2.2: Currently, 4078 protein families exist over all EC classes. Our routine for grouping proteins by EC number and 90% sequence identity is shown schematically below. The dashed arrows represent a protein with two EC numbers being added to two EC classes. The bold arrows show how a protein with no EC number is added to an EC class by sequence identity. The bold arrows represent a protein that is nearly identical to the dashed protein, so it is added to the same two classes. The gray arrow notes that the homologous protein families are compared in the end, and entries found multiple in families are corrected.



1. Use BLASTp to compare each protein chain of each entry to all other chains.

2. All protein sequences are initially grouped into classes by the EC numbers. If a protein has more than one EC number, it is a member of more than one EC class (dashed arrows in Figure 2.2).

3. Structures that do not have an EC number are checked against the existing EC classes. If the sequence is 90% identical to any protein in an EC class, the sequence is added to that class. These entries can be added to more than one class (see bold arrows in Figure 2.2).

4. Any structures that do not have matches in the EC classes are initially grouped into a nonenzyme class. The nonenzyme class can contain enzymes that lack EC numbers or proteins that bind ligands but do not catalyze a reaction.

5. Homologous protein families in each EC class are created using the comparison matrix generated from step 1. At this stage, two entries (A and B in a class) are grouped together into a homologous family if one of the sequences in A is 90% identical to one of the sequences in B. With 90% sequence identity being so strict for clustering, we always found that any additional chains in entries A and B were also 90% sequence identical.

6. In some cases, every entry in an EC class may be at least 90% identical to all other entries. In those cases, the entire EC class is grouped into one homologous protein family. In the nonenzyme class, there are many, different homologous protein families because of the greater structural diversity.

7. At this point, the homologous families within all EC classes are compared to identify any potential errors.

    (a) For proteins with more than one EC number, we find nearly identical protein families in more than one EC class. Only one of the families is retained and placed in the most appropriate EC class.

    (b) If an error was made in the EC number of an entry, it will initially be placed into the wrong EC class, but it will have little similarity to the other entries in that class. The misplaced entry will have high similarity to the entries in another protein family in the correct EC class (e.g., HIV protease was given many different EC numbers for historical reasons, but the entries must be grouped together). The incorrectly labeled entry is moved to the proper class/family. At this time, a missing or incorrect EC number in Binding MOAD can only be corrected if the entry can be

identified by its similarity to a homologous protein family in the proper EC class.

8. The best entry in a protein family is the structure with the tightest binder. In cases where a family has no entry with binding data, complexes of ligand-protein or ligand-cofactor-protein are chosen over protein-cofactor complexes. The priority for choosing a representative of the protein family is:

   (a) Tightest binder (when binding data available)

   (b) Best resolution (complexes with ligands preferred over complexes with just cofactors)

   (c) Wild-type over structures with site mutations

   (d) Most recent deposition date

   (e) When all criteria are the same, the representative is chosen based on comments in the crystallography paper.

2.2.5 Annual Updates

We conduct updates annually to incorporate more structures into Binding MOAD as they become available in the PDB. Our 2004 update began in August. The update procedure is:

1. Use the PDB's list of obsolete entries to identify any existing structures in Binding MOAD that should be removed.

2. Download a new set of mmCIF files. The previous version will be compared to identify all new structures that have been added to the PDB since the last version of Binding MOAD was created.

3. Identify good protein-ligand complexes in the new structures using our current scripts.

4. Any new HETs must be classified as suitable ligands or added to the suspect, partial, or reject lists.

5. The literature portion of the updates should be faster because the number of complexes will be significantly smaller than the existing set and almost all references will be available as online PDF files.

6. Sequences will be added to existing classes and protein families, but regrouping all sequences from scratch may be necessary to periodically confirm our protein classes and families.

7. Each new structure will be compared with the leader of its homologous protein family to determine if the new structure is a better representative of the family.

2.3 Results and Discussion

The creation of Binding MOAD has been the compilation of many years of work and has had several people assist with the project. I am directly responsible for writing the perl program used to filter the Protein Data Bank. The description of the filtering of the PDB and the generation of the data in binding MOAD has been placed in Appendix C. I was also responsible for intitiating the use of the list of ligands that are to be considered "suspect" ligands and to be investigated by hand. Upon going through each 'HET' group by hand to determine whether it is a valid or suspect ligand, another class of ligands was determined, the "partial" ligand list from Table 2.1. I was also involved with the decision of how to handle metals when determining whether a ligand was covalent.

The undertaking to go through the thousands of literature citations to pull out binding affinity values was shared among several people, with the majority falling on my shoulders. In the first two updates, the papers were viewed in '.pdf' format in Adobe Acrobat Reader. In order to make the process more manageable, I used

keywords to search the paper, such as 'ki', 'kd', 'ka', 'ic50', 'affinity', 'bind(ing)', 'constant', 'association', 'dissociation', 'inhibitor', and 'inhibition'. These keywords as well as other combinations were also used to build a dictionary of terms that was used in BUDA, which is discussed further later in the chapter.

After determining the complete set of Binding MOAD, the entries were grouped by family, with enzymes annotated with the EC number. I used several common keywords for entries that were not provided EC numbers to sort them into the categories listed in Table 2.2.

After examining the PDB contents in our latest updated, January 1st, 2009 (55,072 entries), a total of 13,138 valid protein-ligand complexes was obtained. Table 2.2 provides detailed information about the functional roles of the proteins contained in Binding MOAD. Our distribution of structures is a little different than that of sc-PDB(123) due to slightly different selection criteria. Three-fourths of the proteins are enzymes, with hydrolases and transferases having the most representatives.

Binding MOAD contains 6213 unique, valid ligands within the 13,138 complexes. Cofactors, inhibitors, and substrates are all considered ligands in Binding MOAD. Figure 2.3 provides the distribution of valid ligands by size. The ligands range from 4-176 heavy atoms. The average molecular weight of the ligands in Binding MOAD is 455 g/mol; an example of the average ligand is ATP which has molecular weight of 507 g/mol. Figure 2.3 shows that the number of significantly larger ligands drops off quickly. The largest ligands are peptide, nucleic acid, and sugar chains.

2.3.1 Clustering Binding MOAD into Homologous Protein Families

The protein sequences of the entries in Binding MOAD were grouped into homologous protein families. When the set is clustered at 100% sequence identity, 7247 unique protein sequences were identified. As one would expect when the criterion for sequence identity is relaxed, fewer protein families are found and the size of the

Table 2.2: Functional classification of current entries in Binding MOAD

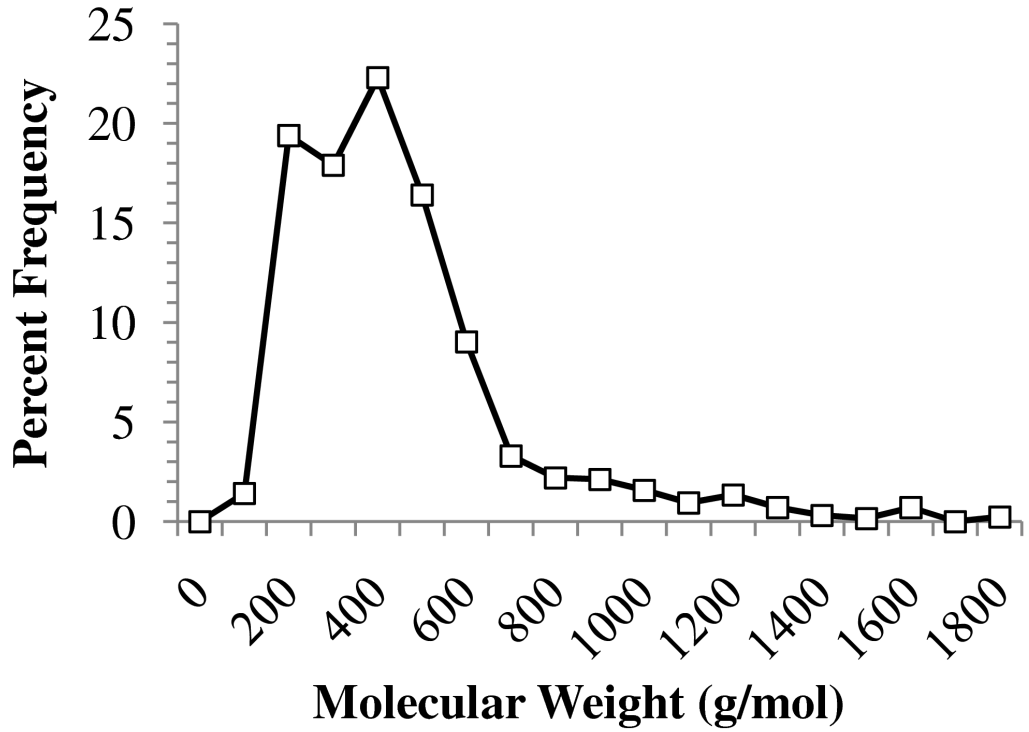
Proteins identified with EC numbers[a]	**Entries**[b]
1.-.-.- (OXIDOREDUCTASE)	2168 (16.5%)
2.-.-.- (TRANSFERASE)	2927 (22.3%)
3.-.-.- (HYDROLASE)	3641 (27.7%)
4.-.-.- (LYASE)	723 (5.5%)
5.-.-.- (ISOMERASE)	463 (3.5%)
6.-.-.- (LIGASE)	331 (2.5%)
Total enzymes	10253 (78.0%)

Proteins without EC numbers	**Entries**
Binding (lectin, streptavidin, agglutinins, etc.)	593 (4.5%)
Signalling, cell cycle, apoptosis	450 (3.4%)
Folding (chaperones, etc.)	67 (0.5%)
Immune (antibodies, immunoglobulins, cytokines, etc.)	294 (2.2%)
Mobility/structural (actin, myosin, etc.)	94 (0.7%)
Toxin/Viral	87 (0.7%)
Transcription, translation, replication proteins	320 (2.4%)
Transport (amino acid transporters, electron transport, etc.)	414 (3.2%)
Enzymes without EC numbers (eg., isopenicillin N synthase)	83 (0.6%)
Other	483 (3.7%)
Total proteins without EC numbers	2885 (22.0%)

[a]Enzyme counts include entries without EC numbers that could be identified through keywords or enzyme names. Some were also identified by 90% sequence identity to entries with EC numbers.

[b]Number of entries and their percentage of all 11,368 entries in Binding MOAD

Figure 2.3: Distribution of the current 6213 unique ligands by molecular weight. The average ligand in Binding MOAD is 455 g/mol. The largest are small chains of sugars, amino acids, and nucleic acids.



protein families increases (Table 2.3). Clustering at 90% sequence identity (our preference) produces 4078 homologous protein families with the largest family containing 278 complexes. The largest families are for systems that have been well studied for molecular recognition between proteins and ligands (e.g., trypsin, thrombin, HIV protease, lysozyme, dihydrofolate reductase, etc.). In Figure 2.4, a histogram of the homologous protein families shows that most of the families have only a few entries. This reflects the emphasis in structural biology to identify new structures and folds, rather than solve many structures of the same protein. Generally, families contain multiple complexes when mutagenesis studies have been performed or various ligands have been co-crystallized.

Table 2.3: Characteristics of Binding MOAD When Grouped Into Families by Sequence Identity

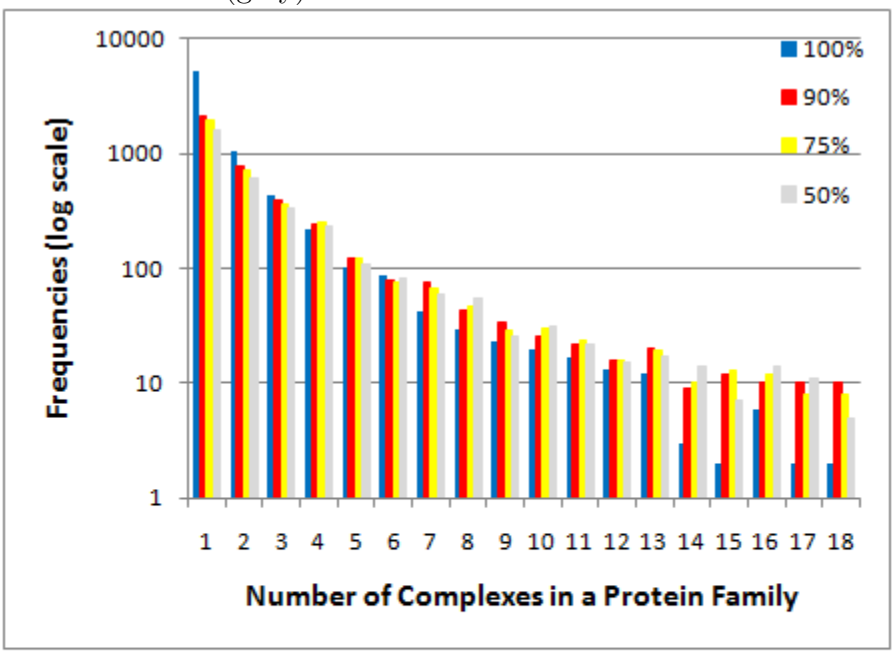
Clustering Criterion	Number of homologous protein families	Size of the largest family (second largest family is also noted)
100% Sequence identity	7247	124 complexes[1] (52)[2]
90% Sequence identity	4078	278 complexes[3] (165)[1]
75% Sequence identity	3823	272 complexes[3] (182)[1]
50% Sequence identity	3316	272 complexes[3] (190)[1]

[1]Trypsin
[2]Thrombin
[3]HIV Protease

2.3.2 Nonredundant Binding MOAD

To create a nonredundant version of the dataset, we had to choose unique representatives for each protein family. As outlined in the Methods, we made every effort to identify the tightest binder to represent each family. For the dataset clustered at 90% sequence identity, 2107 of the 4078 families contained only one complex, and so the choice for the representative was obvious. The remaining families contained multiple complexes. In the 2008 update of Binding MOAD, for 724 of the families, the representative was easily identified by binding data. Resolution was the deciding

Figure 2.4: Histogram of the homologous protein families shows that most families have only a few complexes. There is a near-exponential decrease in the number of larger and larger families. This trend is basically the same for clustering at 100% sequence identity (blue), 90% (red), 75% (yellow), and 50% (gray).



factor for 335 of the families (either because there was no binding data or the binding affinity was the same for more than one ligand). Of the remaining families, 46 were chosen based on complexes with ligands being preferred to complexes with only cofactors, 13 were chosen by wild-type over mutated protein, 24 by most recent deposition date, and 48 by other criteria (R factor, comments about ligands in the paper, etc.). In the current version, the new structures were checked against the previous leader to determine the leader.

The nonredundant version of Binding MOAD contains 4078 unique proteins. After choosing the complexes for the nonredundant set as outlined above, this set contains binding data for 1176 of the unique structures.

2.3.3 Binding-Affinity Data

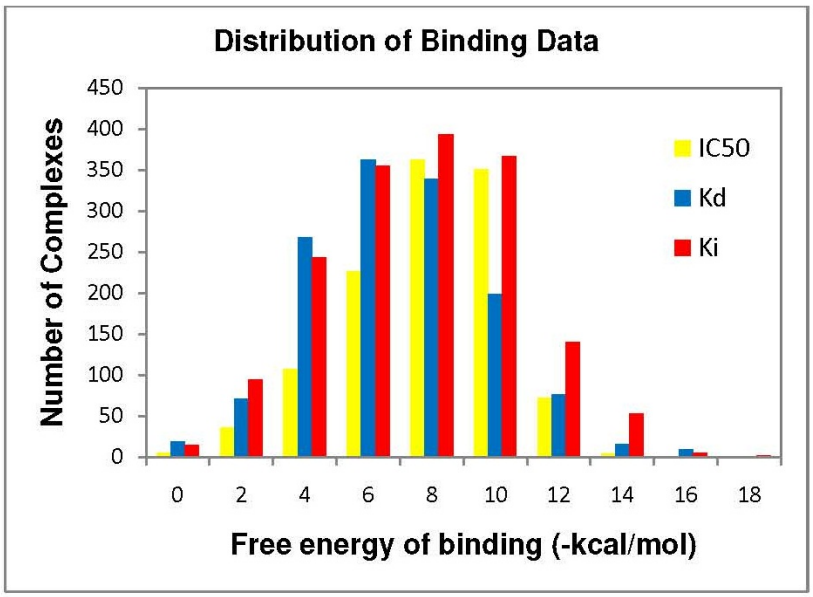
The binding-affinity data contained within Binding MOAD ranges 13 orders of magnitude, from low fM to high mM values (see Figure 2.5). The dataset contains mostly K_d and K_i values. There are 1167 entries have IC_{50} data, ranging 21 pM - 125 mM. For the 1365 entries with K_d data, values range 77 fM - 900 mM. The 1671 entries with K_i data have the largest range of binding affinity, 11 fM - 400 mM.

One of our primary goals is to obtain binding data for all entries in the full set of Binding MOAD (all 13,138 complexes). At this time, only 4203 complexes (30%) in Binding MOAD are augmented with binding data. Though this is much larger than other datasets with a few hundred binding affinities,(54; 55; 114) we were disappointed to find that so few of the structure papers notes binding-affinity data. A survey of the literature by Wang and coworkers found a similar rate of binding data included in the crystallography papers.(56)

Of course, some of our complexes inherently lack binding data; protein-cofactor structures do not have K_d, K_i, or IC_{50} data for us to report. K_M is the more appropriate binding data for most cofactor-protein complexes, and we have started to

collect that information for our complexes. Protein-cofactor structures should be part of the dataset because they can be very important in studying molecular recognition and drug design. For example, patterns in ATP recognition can be extracted from ATP-binding domains to explain enzymatic regulation or develop inhibitors.(131; 132)

Figure 2.5: The distribution of binding-affinity data within Binding MOAD. Data is available as K_d (red), K_i (blue), or IC_{50} (yellow). For this histogram, binding data were converted to free energies by -RT ln (data). Though not strictly appropriate for many K_i or IC_{50}, this simply provides a comparison for the reader.



2.3.4 Database Growth and Updates

As mentioned above, we are committed to the growth and quality of Binding MOAD. Since its introduction in 2004, Binding MOAD has regularly expanded its collection with new data. Originally with 5331 crystal structures of protein-ligand complexes, it has increased by almost 1500 each year, growing to 6638 in 2005 and then 8250 in 2006, reaching 9836 entries in 2007, 11,368 in 2008, and 13,138 with the latest update. This steady growth mirrors the growth of the PDB (Binding MOAD contains approximately one-fourth of the PDB). The primary literature for each crys-

tal structure is read in order to verify the ligand and to extract any affinity data for the ligand. Thus, adding new data to Binding MOAD involves reading tremendous number of journal articles for manual annotation and validation of appropriate ligands.

To facilitate the literature-checking process, a natural language processing (NLP) based workflow tool called Binding Unstructured Data Analysis (BUDA) has been developed. The NLP portion of BUDA is built upon the General Architecture for Text Engineering (GATE) framework(133). It identifies key sentences and phrases in papers and uses a weighted scoring algorithm to rank the likelihood that the key sentences and phrases contain binding data. The workflow portion of BUDA is used to interact with the researcher to organize the data for the annotation process. From the workflow interface, the curators can sort the articles by their weighted scores, review the annotated texts and highlighted sentences, and update the data into Binding MOAD.

2.3.4.1 Platform

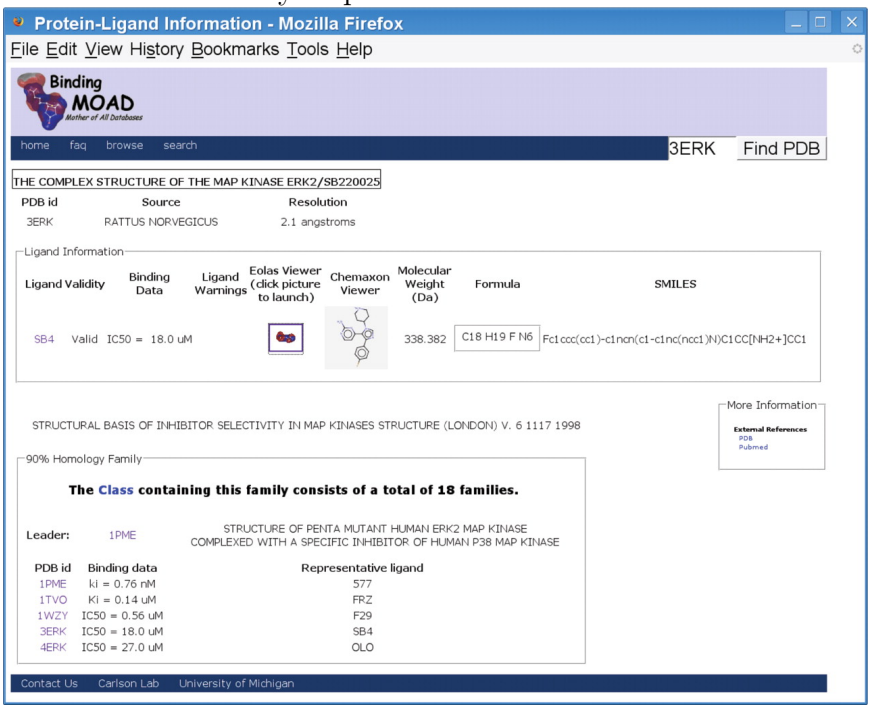
Binding MOAD is built on proven technologies. The Binding MOAD database is based on the Java 2 Platform, Enterprise Edition (J2EE), using an open-source JBoss Application Server, Enterprise JavaBeans (EJB), and a MySQL database backend. These tools provide a standards-compliant, easy-to-use website that unifies the presentation of structural, chemical, and binding data in one simple format.

2.3.4.2 Improving User Experience

Having a flexible infrastructure, allows for changes in the web-site presentation. Efforts are made to make the data as easily accessible as possible. We have removed the need for users to login, and data is now freely accessible to private companies, non-profits, and foreign institutions. Additional features have been added. A screenshot

of the modified layout for a datapage in Binding MOAD is shown in Figure 2.6.

Figure 2.6: Screenshot of the data page for 3ERK, showing the additional ligand data and the connectivity to proteins with similar structure and function.



2.3.4.3 Viewer

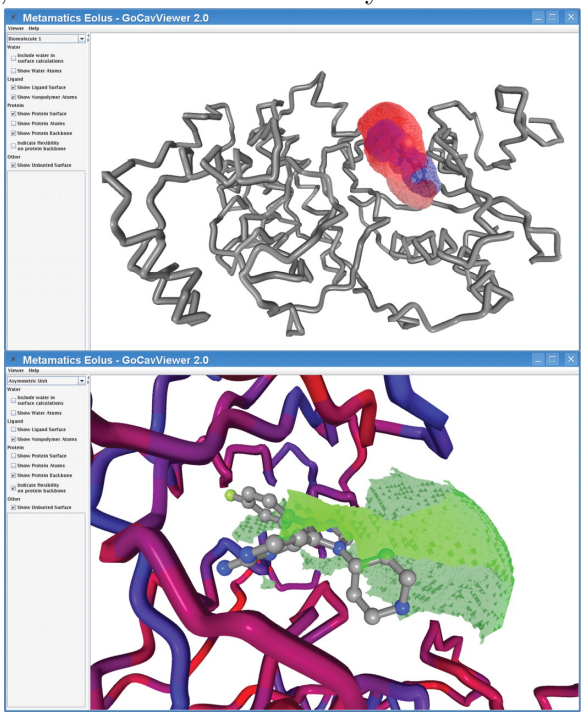
A new 3D protein viewer, EolasViewer, is available to view the ligand in the protein pocket. The new viewer is built using the Eolus platform from Metamatics and it replaces the previously used GoCavViewer. A screenshot of the viewer is shown in Figure 2.7 The new viewer is still capable of selecting and viewing the ligand pocket using both ball-stick and surface representations. EolasViewer incorporates significant improvements in the areas of performance, visual quality, and back-end flexibility for future application development efforts.

By taking advantage of rendering algorithms and OpenGL Shader Language (GLSL), Eolus provides the new viewer with new representation styles. The surface representation has been expanded to a fully transparent polygon surface. The proteins are rendered as ribbons by default, and the entire protein (instead of only the ligand

pocket atoms) can now be rendered either as ribbon or ball-stick. Finally, many advanced features are planned for future versions of this tool. Eolus is a platform for structual biology being developed in conjunction with this and other tools.

Like its predecessor, the new Eolus-based viewer is built using a Java framework and we are deploying it as a WebStart application. Eolus uses Jogl (Java Bindings for OpenGL) to fully utilize the 3d acceleration features available in nearly all modern computers. These two technologies, Java WebStart and OpenGL, provide nearly hands-free deployment of the software, together with state-of-the-art performance and visual quality.

Figure 2.7: EolasViewer for 3ERK. The SB4 ligand is shown in ball in stick inside the pocket. The surfaces shown are the ligand surface in blue, the binding site in red and the solvent-exposed regions of the binding site are in green. (Top) The protein backbone is shown as a gray ribbon, and in the close-up (Bottom), the backbone is colored by B-factors.



2.4 Conclusion

As stated above, we have developed and continue to expand Binding MOAD, in the future we wish to contain more binding-affinity data (including the addition of K_M for cofactors). We have also committed to annual updates of the dataset to keep pace with the growth in the PDB. Binding MOAD has over eleven thousand, hand-curated, protein-crystal structures that contain biologically relevant ligands. Binding affinity data is available for almost one-third of the entries. Part of the value of Binding MOAD is in its careful curating and in its size and wealth of data. This has been only achievable because of the efforts invested to maintain the continual growth. Binding MOAD has plans for even greater improvement. We are planning to add similarity-based searches for the ligands. Furthermore, while we have been able to use text-mining tools to speed up our annotation process, we are looking to make these tools available online to allow users to mine text for additional types of data. We are now using NLP to aid in our searching. Such NLP based text mining approaches can be readily applied to other bioinformatic projects. This technology can be used to extract a wide variety of data - not just binding information - from the huge body of literature available today. NLP is proving to be a valuable tool in aiding the curation of Binding MOAD. It has significantly sped up the process of the annual updates of adding data.

We have made the dataset available online at http://www.BindingMOAD.org. This web-accessible resource makes our information freely available to other research groups at non-profit organizations (annual licenses are available to the private sector). Data from our perl scripts and our hand curation include PDB id, EC class, homologous protein family, binding-affinity data, and classification of each ligand in the entry (valid versus invalid). The datapage for each complex in Binding MOAD provides this information to the user. Our scripts also note the reason any PDB structure was excluded (resolution $>$ 2.5 Å, no appropriate ligand, etc.). If a user

tries to access a PDB entry that is not part of Binding MOAD, a datapage provides the reason for its exclusion from the dataset.

We are choosing to make the structures available as biological units rather than PDB files. The biological units provide the proper multimer for biological activity. For instance, only the proper dimer is provided when multiple dimers occupy a unit cell, or the proper tetramer is provided from symmetry operations of a unit cell containing only the monomer. This will provide users with the structures that are most related to biological activity and therefore the most appropriate for study.

CHAPTER III

Exploring Protein-Ligand Recognition with Binding MOAD

3.1 Introduction

A growing trend in computational biology is the development of large datasets to provide the scientific community with various information on protein-ligand structures. Of course, the definitive online resource for structural data of these complexes is the Protein Data Bank (PDB) (53). It is constantly being improved through the addition of online tools and links to complementary datasets (134). Most recently, Ligand Depot was created by the curators of the PDB to facilitate searching the HET groups via chemical substructures and text-based searches (117). There are many other examples of databases and websites that analyze and augment protein-ligand complexes from the PDB. The following discussion is by no means an exhaustive listing of such derivatives of the PDB.

Our own contribution in this area is Binding MOAD (Mother of All Databases) (62). Our goal for Binding MOAD is to create the largest resource of high-quality protein-ligand complexes and augment those structures with binding affinity data and online analytical tools. We took a top-down approach to create Binding MOAD. Starting with the entire PDB, we selected only crystal structures of high resolution

(= 2.5). We ensured that Binding MOAD contained only appropriate protein-ligand structures through extensive hand curation. Structures were required to contain at least one valid, non-covalently bound ligand. Chains of 4 nucleic acids or less and 10 amino acids or less were treated as ligands. In our original creation of the 2003 version of Binding MOAD, we eliminated any structure with a heme group because of the difficulty in distinguishing non-covalently bound ligands. With the August 2004 update, all heme-containing proteins have been examined by hand and appropriate structures are now part of Binding MOAD. The 2004 version of Binding MOAD contains 6821 complexes. We read over 6000 crystallography papers to confirm the validity of the protein-ligand complexes and to gather binding affinity data. As a result of this process, we have binding data for 1793 (27%) of the complexes. The 2008 version now contains 13,138 complexes with binding affinity for 4203 (32%).

We wanted to mine Binding MOAD to provide general patterns of molecular recognition to the scientific community. How exposed binding sites are across all protein-ligand complexes? To answer this, we needed a resource that could properly treat any binding site – regardless of size, shape, degree of solvent exposure, the inclusion of bridging water molecules, or the occurrence of side chains with multiple resolved orientations (partial atom occupancy). For this, we have developed GoCAV and the GoCAVviewer to calculate and display molecular surfaces for the ligands and for the protein cavities.

A number of online tools are already available to view atomic coordinates, secondary structure, and cavities. We are not presenting GoCAV as a breakthrough to supercede these programs. We simply feel that GoCAV and the GoCAVviewer are complementary alternatives to these other excellent resources, and by incorporating the viewer into our website (www.BindingMOAD.org), we have a means to share the data from this study with the scientific community. The discussion below highlights some of the most useful online resources created by other research groups for an-

alyzing and viewing protein-ligand complexes. Generally, these databases describe a ligand as any molecule that is not one of the common 20 amino acids or 8 common nucleic acids. They make no distinction between valid and invalid ligands like crystallographic additives or covalent modifications to the protein.

PDBsum is the most comprehensive resource. It provides data on the entire collection of structures from the PDB (120; 135; 118; 119). Chemical, enzymatic, and genomic information is available for all PDB structures, even if they are not proteins and even if they do not contain ligands. One of the most powerful features of PDBsum for understanding the molecular recognition of ligands is its analysis of macromolecule-ligand interactions. The information is provided via 2D pictures and several 3D viewers. Most relevant to this work is the fact that PDBsum provides analysis of potential cavities using an updated version of SURFNET (48).

CASTp is an online database that uses rigorous analytical techniques to analyze all proteins in the PDB for interior cavity voids and surface pockets (136). Using the program CAST (17), it calculates the volumes and surface areas of the sites, and it also determines the size of the openings in solvent-exposed pockets. It is not limited to proteins with bound ligands, so it has the benefit of identifying previously unknown binding sites, but it also identifies many small surface pockets that do not bind ligands. The online viewer displays the residues that make up the cavities and pockets, but it does not show the bound ligands. This makes it difficult to understand the molecular recognition that controls binding in that site.

MSDsite (58) provides information on ligand interactions with any macromolecule, not just proteins. MSDsite provides various analyses of the macromolecular environment surrounding ligands. The dataset can be mined by matching patterns based on the ligand or on the binding-site environment. PDB-Ligand (137) is a new resource that is very similar to MSDsite, but strictly focuses on analyses of protein residues and nucleic acids within 6.5 Å of a HET group. Relibase (59; 138) is a

resource that specifically focuses on the protein-ligand complexes in the PDB. It allows for text-based and sequence-based searching of the PDB. SMILES strings can be used to search ligand substructures. It also provides graphics tools to examine the structures. Relibase+ (60) is a newer version that allows for additional 2D and 3D similarity searches. NCBI's Entrez resource for 3D structures is the Molecular Modeling DataBase (MMDB) (121). MMDB is based on pregenerated relationships, found by comparing each PDB entry with various structure and sequence databases. Their viewer can be used to compare any individual PDB entry to its structural homologs. This reveals their similar tertiary structure and can be used to examine common binding motifs of bound ligands. So though the focus of MMDB is the comparison of folds and domains, it can provide valuable information on protein-ligand recognition. Each of the four online databases mentioned above has very useful features, but as mentioned above, they make no distinction of which HET groups are proper ligands.

Two additional datasets, PDBbind and sc-PDB, are similar to Binding MOAD and also focus on valid ligands. These databases do not provide viewers to examine protein-ligand complementarity, but the atomic coordinates of the proteins and ligands are available for download and can be examined offline. PDBbind is a large set of protein-ligand complexes from the PDB, focusing on binary structures with a single ligand in a protein binding site (56). PDBbind also provides binding affinity data obtained from reading the crystallography papers. As of its latest update in January 2004, it contains binding data on 1622 complexes (a subset of 900 complexes makes up the "refined" set) (57). PDBbind provides graphical interfaces, similar to those used with Ligand Depot, to view the ligands and perform substructure searches to find related systems. The other database, sc-PDB (123), was created in a fashion similar to Binding MOAD and PDBbind, but it does not provide binding data. The set of structures is used for "inverse screening," a procedure where a ligand is docked to a series of binding sites to determine its appropriate target. sc-PDB is a set of

5634 protein binding sites and 7109 ligands at the time of writing this paper [personal communication, Esther Kellenberger, Universit Louis Pasteur, placeCityStrasbourg]. The online interface to the dataset allows for text-based searches of much of the information within the PDB files (PDB ID, HET group name, authors, EC numbers, deposition date, resolution, etc.). The data can also be accessed by information based on other resources like Swiss-Prot (139) data and NCBI taxonomy notation (140).

3.2 Methods

Rather than simple PDB files, "corrected biounit files" were used for all protein-ligand complexes. Biounit files are available from the PDB, and they represent the appropriate multimer for biological activity. For instance, if only a monomer appears in the unit cell, but a trimer is the appropriate biounit, the other two monomers are generated through symmetry operations. We found that HET groups and water molecules frequently were not properly treated in the PDB's biounit files. They were not propagated where necessary, and they were not removed in cases where their corresponding protein was deleted from the unit cell. We corrected all biounit files by propagating the water and ligands as necessary using the program PyMOL (141). We also removed any molecules that were more than 10 Å away from the protein. Covalent links were checked to avoid truncating sugar chains and other post-transcriptional modifications that were longer than 10 Å. These corrected biounit files are the structures that are available for download on the Binding MOAD website.

It is straightforward to calculate the surface of an enclosed binding site (142; 143), but many interesting ligands are bound in open clefts. Molecular surface area (MSA) is calculated by "rolling a solvent probe" on the van der Waals (vdw) surface of the atoms. With an exposed binding site, the probe escapes and maps out the entire protein surface. Ho and Marshall suggested that some cutoff distance to the ligand might be a reasonable way to determine the boundary of an open site (144). We were

not able to find code to do this, so we wrote GoCAV to accomplish the task and provide a consistent treatment for any type of binding site in Binding MOAD.

GoCAV, uses an "enlarged ligand surface" (ELS) to create a boundary for the binding site (Figure 3.1). It calculates MSAs using a grid-based method. Voronoi tessellations are more accurate than grids for enclosed sites (145; 146; 143), but the method does not work as well on surfaces (147; 148). We use a very fine, 0.2-Å grid (0.008 $Å^3$ cubes) to minimize the errors as much as possible. Codes have been developed by other groups that calculate surfaces and cavities (for example, POCKET (47), SURFNET (48), CAST (17), PASS (49), and an unnamed grid-based technique by Schneider and coworkers (149)). Many of these have the benefit of finding pockets without needing bound ligands to guide them, which means they can identify new binding sites (a definite advantage over GoCAV). However, in the process of analyzing/identifying all possible cavities, some of these codes produce pockets that are not true binding sites. Some have poorly defined boundaries that do not encapsulate all of a bound ligand. Some do not identify all types of pockets, and others tend to create large networks of interconnected cavity spaces over the surface of the protein. For our purposes with Binding MOAD, we needed a code which focuses on defining a cavity within the local vicinity of a bound ligand.

To create the ELS, we extended the ligand's vdw radii by 2.8 $\mathring{A}$ (the equivalent of one layer of water as the exterior boundary for an open binding site). We wanted to define a boundary for open binding sites, but not hinder the calculation of enclosed binding sites that incorporate bridging water molecules. To verify our description of the ELS, we examined several binding sites with bridging water molecules (see Figure 3.2). Appropriate boundaries of these binding sites were identified with the ELS radius of 2.8 $\mathring{A}$. The MSA for ligands are straightforward to calculate and were also part of the GoCAV output.

The overwhelming majority of Binding MOAD's structures do not contain hydro-

Figure 3.1: Determining the boundary of an open cavity using ELS. (Left) A ligand molecule (black) is bound in an open protein cleft (gray). The dashed line is the ELS, determined by adding 2.8 $\mathring{A}$ to the radii. A probe rolls over the vdw surfaces of the protein atoms and the inward-facing surface of the ELS. The resulting surface of the cavity is shown as a bold, black line. The solvent-exposed portion of the cavity surface is defined as the section of bold, black line that is defined only by the ELS in the opening of the binding site.

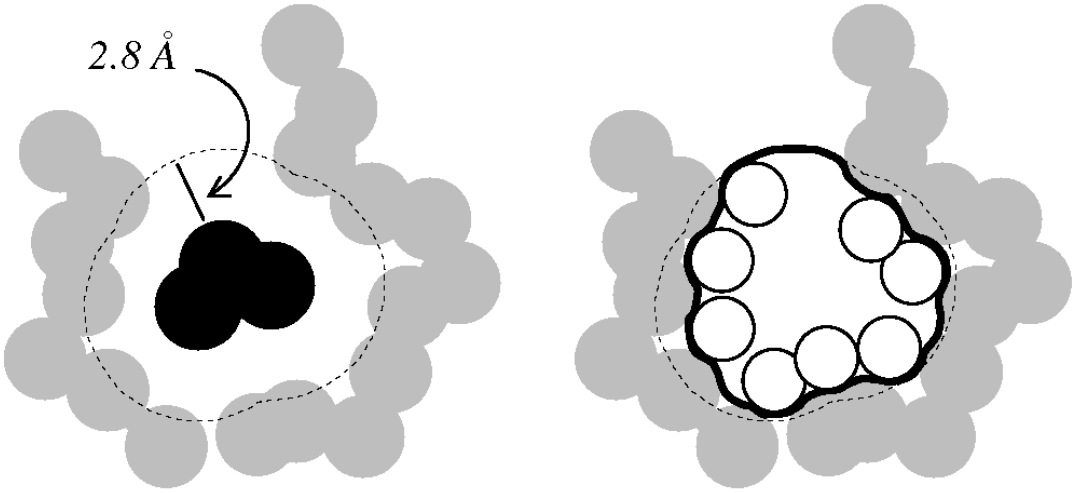
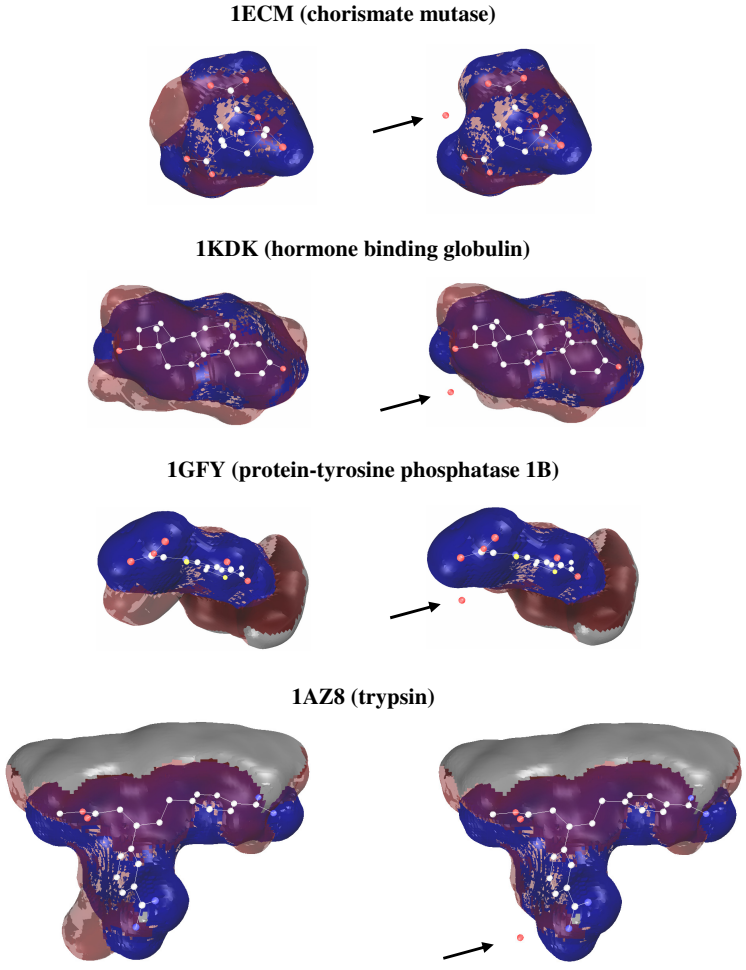


Figure 3.2: The use of an ELS does not create inappropriate boundaries for open or closed cavities that contain bridging water molecules. Examples are given for completely buried cavities (1ECM and 1KDK) and solvent-exposed pockets (1AZ8 and 1GFY). (left) Binding site and ligand surfaces calculated with GoCAV, employing an ELS cutoff. (right) The resulting surfaces when the noted bridging water molecules within the cavity are included in the calculation as additional protein atoms. The ligand surface is blue, and the binding site surface is red and gray. The red regions are buried, and the gray region denotes the solvent-exposed or ELS surface of the cavity. Protein atoms are not shown for clarity. This figure was created using the GoCAVviewer on the Binding MOAD website.



gen atoms, so we needed to use united-atom radii in our analyses. Our chosen radii were based on averaged OPLS united-atom vdw parameters: C=1.925 Å, N=1.655 Å, O=1.52 Å, S=1.81 Å, P=1.87 Å (150). (We estimated radii for other less-typical atoms as 2.0 Å.) OPLS parameters were carefully developed to reproduce thermodynamic properties in condensed phases. We are confident in the choice of OPLS radii because of their good agreement with Li and Nussinov's radii set which was determined in an entirely different fashion (151). Li and Nussinov derived radii through contact distance distributions in a set of 1405 protein crystal structures (C=1.92 Å, N=1.66 Å, O=1.51 Å, S= 1.92 Å). Though we do not present the data here, a user can use a second set of radii in GoCAV. We made the Fleming and Richards' radii (147) (C=1.9 Å, N=1.5 Å, O =1.4 Å, S=1.85 Å) an available option because they are well established and many groups support smaller radii. Gerstein and coworkers determined similar, smaller radii using contact distance distributions from crystal structures of small organic molecules (C=1.88 Å, N=1.64 Å, O=1.44 Å, S= 1.77 Å) (152).

Invalid ligands are not included in the calculations unless they are a covalent modification of the protein or a structural element like a catalytic/structural zinc ion or a heme (these are treated as additional protein atoms). When mining a large dataset, the code must properly treat unusual cases. GoCAV was created with several "filters" to analyze structures before performing the surface calculations. With these filters, GoCAV was able to properly process >98% of the structures in Binding MOAD. In the case of ligands with warnings (too many or too few atoms), those complexes were not included in this study (even when GoCAV was able to calculate their surfaces).

Unusual protein-ligand complexes include the following situations: 1) Side chains within a binding site can be solved in multiple orientations (as denoted by partial occupancy). In these cases, GoCAV automatically calculates the surfaces twice, with the side chain in either orientation. Appropriate combinations are generated if more than

one side chain in the binding site requires this treatment. All solutions are presented in our analysis, providing averaged data points with error bars for those complexes. 2) Some sugar-binding proteins actually contain both enantiomers of the sugar in the binding site, superimposed with 50%-50% occupancy. Again, GoCAV recognizes the two solutions inherent in the structure and does two independent calculations, each with a single enantiomer. Both ligands are presented independently in our plots and histograms (no error bars because they are not the same ligand). 3) When two separate ligands are accommodated in a large binding site (such as a cofactor and an inhibitor bound in close proximity), GoCAV actually does three calculations: a) both ligands are treated as one large molecule, b) the first ligand is treated as part of the protein while the surfaces around the second ligand are calculated, and c) the second ligand is part of the protein while the first ligand is calculated independently. The later two calculations, where each ligand is treated independently, are the values included in the plots and histograms in this study.

To verify that the patterns calculated with GoCAV are appropriate and comparable to other standard techniques, we have also calculated the solvent accessible surface area (SASA) of the ligands with the program NACCESS (50). (SASA of the ligand should be roughly comparable to the MSA of the cavity.) NACCESS is based on Lee and Richard's analytical method (51) as opposed to our grid-based approach. It uses radii based on Chothia's (153) but with more subtypes for carbon, nitrogen, and oxygen with slightly different radii (carbons range 1.76-2.0 Å, nitrogens range 1.5-1.65 , oxygens are 1.35 or 1.4 Å, S=1.85 Å, P=1.9 Å, Fe=1.47 Å). NACCESS provides SASA on a "per residue" basis. If the ligand is completely buried, the SASA calculated with NACCESS is zero. The SASA of each ligand in each complex was calculated in the presence and absence of the protein (again, we included any other appropriate ligands as part of the protein environment). We calculated the buried surface area of the ligand as SASA(no protein)-SASA(with protein) and percent buried

surface area as 100*(1-SASA(with protein)/SASA(no protein)). NACCESS is not able to treat the unusual cases that we describe above for GoCAV. Those systems are not included in the NACCESS plots and histograms.

3.3 Results and Discussion

3.3.1 Binding MOAD

Several features of Binding MOAD make it particularly useful for examining the degree of solvent exposure of all protein-ligand binding sites. First, the dataset has been carefully curated to identify valid and invalid ligands in each structure. Only the valid ligands are included in our analysis. Without this analysis, any broad mining of the structures would reflect real binding patterns skewed by the less relevant patterns seen for crystallographic additives. (We have also excluded any ligands with warnings of too many or too few atoms from the analysis, though they are part of the MOAD dataset.)

Second, the dataset has been analyzed for redundancy. The proteins have been grouped into families by 90% sequence identity. The non-redundant set of structures from Binding MOAD contains only one complex from each protein family. The representative for the family is the tightest binder when binding data is known. In cases where there is no binding data for any of the complexes in the family, the representative is chosen based on best resolution and other structural considerations (62). This allows us to present the data without some of the inherent bias of structures deposited within the PDB.

3.3.2 Sharing the data on the Binding MOAD website

Each entry's datapage on the Binding MOAD website is organized to help users identify related protein systems and compare binding data, see Figure 3.3. Entries

are cross-linked by function (classes for both enzymes and non-enzymes), sequence identity, and ligand content. All HET groups in the complex are identified as valid or invalid, and warnings are provided when too few or too many ligand atoms appear in the PDB entry (unresolved atoms or multiple resolved orientations for parts of the ligand, respectively). Binding data is provided when available. Text-based searches can be used to identify entries based on PDB id, EC number, protein name, 3-letter HET codes, and authors. Wildcards are permitted. The results can be limited to a user-defined range of crystallographic resolution. The user can also limit the search to the 1793 structures in Binding MOAD with available binding data, the 2223 structures of the non-redundant Binding MOAD dataset (where each protein family is only represented once), or the 630 structures in the non-redundant set that have binding data. There is also a browse feature to allow users to page through functional classes of structures. When a user clicks the "class" link on a datapage (seen in Figure 3.3), they are taken to the browse page for that functional class where all protein families within the class are shown and ligand/binding information also provided (Figure 3.4).

Clicking the blue and red thumbnail on a datapage, see Figure 3.3, launches a version of the GoCAVviewer that interactively displays the atomic coordinates and the surfaces calculated with GoCAV. The binding-site surfaces and the ligand surfaces calculated with GoCAV are grid points, so the "raw" surfaces look like LEGO building blocks. A smooth surface is created by a graphics trick, applying a Gaussian filter to the image. The GoCAVviewer is written entirely in standards compliant Java, and the code will work on any operating system that provides an implementation of the standard Java Runtime Environment and Java3D API. GoCAVviewer is interactive, allowing the user to rotate, zoom, or translate the structures in real time. The cavity surfaces are transparent and near-by protein atoms can be displayed, so the user can look at the complex in detail. At this time, the most critical issue is speeding up the

Figure 3.3: The datapage for the HIV-1 protease complex 1MTR. The page starts with the general information from the PDB file. The ligand HET codes are single-click searches that pull up all other structures with that ligand. All ligands are listed as valid or invalid, and binding affinity data is provided when available. Warnings are provided when the number of atoms in the structure do not match the formula section of the PDB file. Clicking the thumbnail launches the GoCAVviewer. Links to the right of the thumbnail take the user to the equivalent datapage at the PDB and to the crystallography paper on Pubmed. Various sets of structural and binding data are available for download. At the bottom of the page, the structure is linked to other entries with the same functional class, and all other members of its protein family are listed with ligand information (over 100 HIV-1 protease structures are included in Binding MOAD and the user needs to scroll down the page to see all the data.

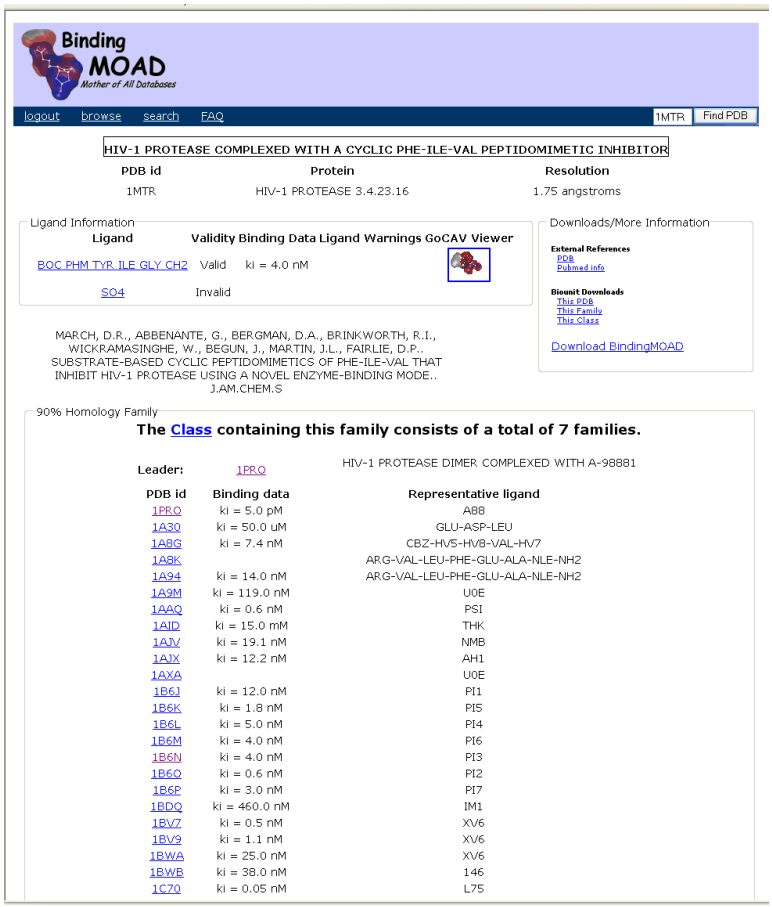
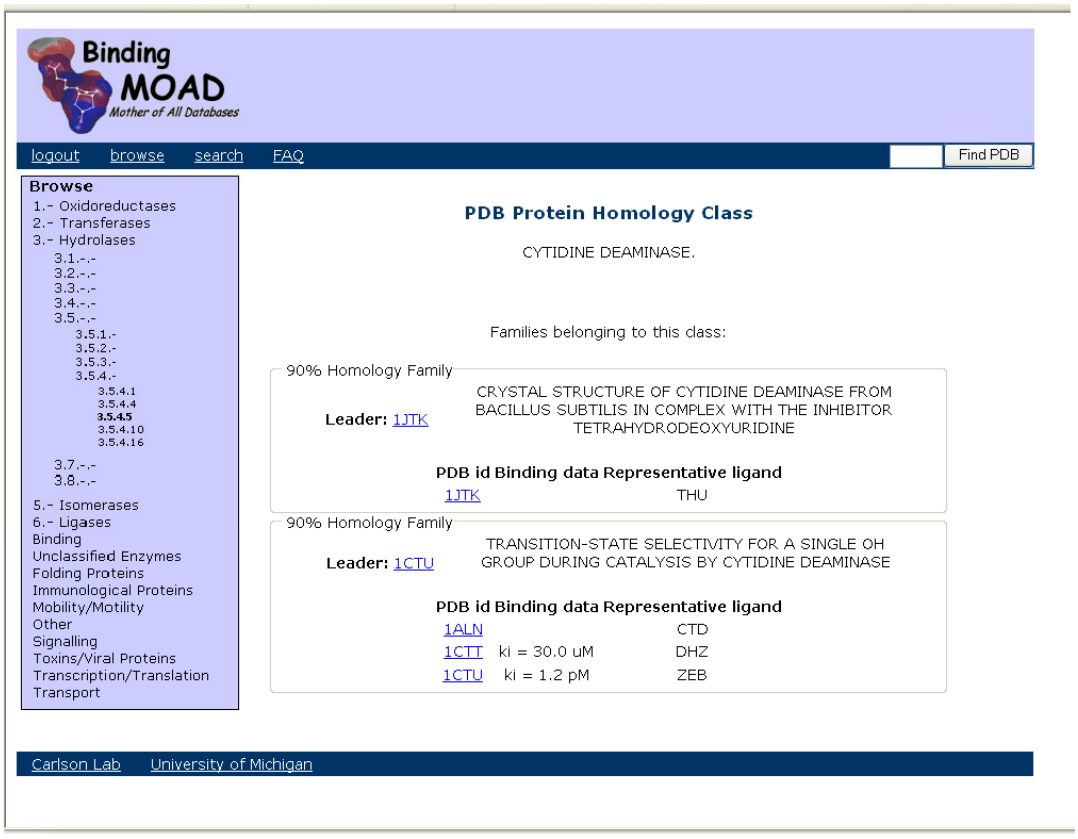


Figure 3.4: The user can find information by browsing through the complexes within Binding MOAD. The structures are organized by function: EC numbers for enzymes and our own classifications for entries without EC numbers. All protein families within a class are displayed for the user to compare related systems and their binding affinity data.



viewer. We are committed to improving it, but we wanted to make the data available to the rest of the community as soon as possible.

Tight complementarity between the protein and ligand is highlighted by the ligand surface projecting through the cavity surface (see Figure 3.2). We have found that these intersections only occur at positions with strong hydrogen bonding or very specific vdw interactions. We have also configured the viewer to display a second set of surface information calculated with bridging water molecules. It was easy to include water molecules as additional protein atoms in a GoCAV calculation and determine their influence on creating a surface to complement the ligand. Figure 3.2 shows how the surfaces change when bridging waters are treated as part of the protein. The shape complementarity between the ligand and the pocket is often more evident when waters are included.

3.3.3 Mining Binding MOAD

Figure 3.5 provides histograms of the size of the ligands in the redundant and non-redundant Binding MOAD sets. The distribution of ligand sizes is similar in the two sets. In Figure 3.6, the plots of size vs. affinity show the wide range of data available in Binding MOAD. One issue that should be noted is that the "affinities" used in our plots are a simplistic translation of the Kd, Ki, and IC50 data using the formula RT*ln(data). This is not strictly correct for Ki or IC50, but it is a way to do a standardized treatment of a large dataset. The Kd data in the plots is highlighted in black because the affinities should be more reliable and better reflect true free energies of binding. Complexes with Ki and IC50 data are in gray. The data available for download from the website is the original Kd, Ki, and IC50 data from the crystallography papers.

The ranges in Figure 3.6 are approximately the same for the redundant and non-redundant sets, but the averages for both sets are slightly different. The average

Figure 3.5: Distribution of ligand size within the complexes in redundant and non-redundant Binding MOAD, note the larger scale for the redundant complexes. Black bars represent all complexes in Binding MOAD; gray bars represent only the complexes with affinity data.

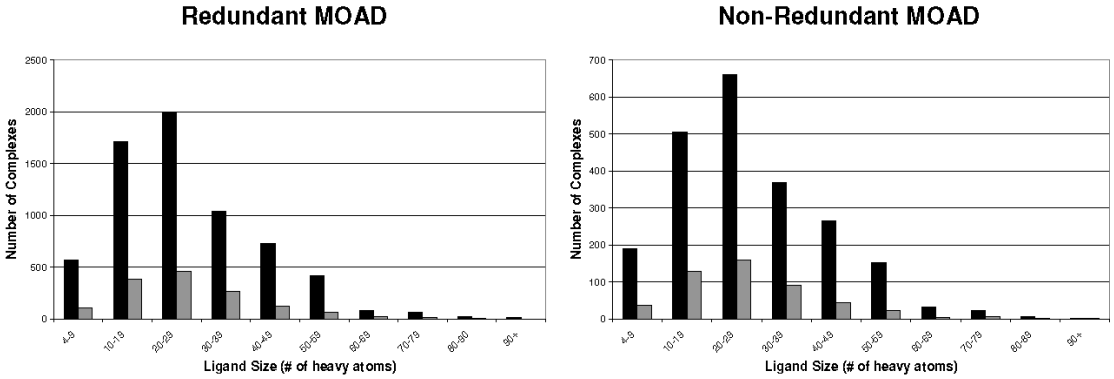
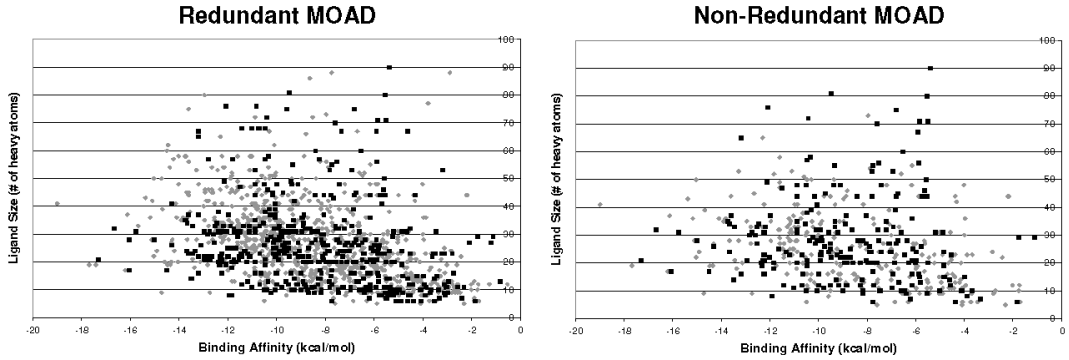


Figure 3.6: Plots of ligand size vs. binding affinity for the complexes in redundant and non-redundant Binding MOAD. The data points in black squares are from complexes with K_d data, and gray diamonds are used for complexes with Ki or IC50 data.

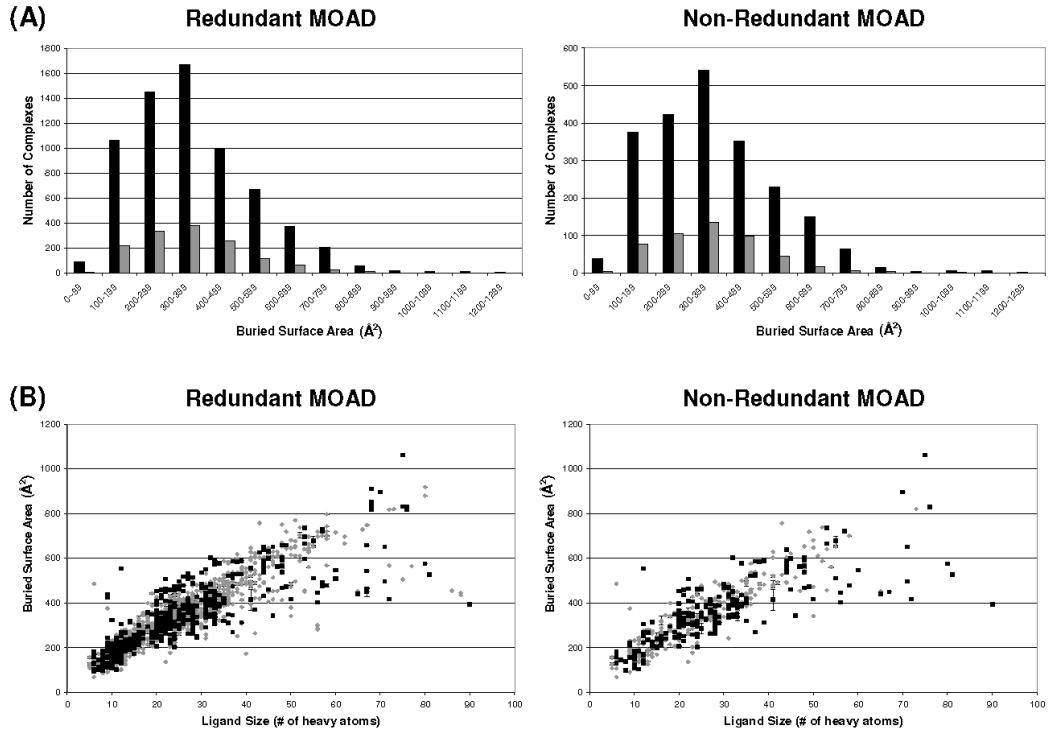


binding affinity for the redundant set is -8 kcal/mol, but the average for the non-redundant set is -9 kcal/mol. Both sets have a standard deviation of 3 kcal/mol. The average numbers of heavy atoms for the ligands in these sets are 26 and 27, respectively, both with a standard deviation of 14 atoms. A size range of 12-41 heavy atoms corresponds to drug-like molecular weights of approximately 150-700. It should be noted that these are the averages for just the complexes with binding affinity data (all points in Figure 3.6, but only the gray bars in Figure 3.5). The average number of heavy atoms for ligands in all of the Binding MOAD complexes is 31 (black bars in Figure 3.5).

Figure 3.7A presents histograms of buried MSA for the binding-site cavities as calculated by GoCAV. The distribution of buried surface area parallels the size distribution of ligands, and in Figure 3.7B, a plot of the cavity's buried MSA vs. ligand size shows a good correlation, simply reflecting the relationship between increasing size of the ligand and increasing surface of the cavity it occupies. (As expected, the distributions for buried SASA of the ligands, as calculated with NACCESS, were very similar and also well correlated to ligand size, data not shown).

Liang et al. found that a linear correlation exists between ligand volume and binding site volume, provided that the pockets were small (=700 $Å^3$). Figure 3.7B also shows that the correlation is not as tight for the larger ligands and pockets. Others have found that binding sites tend to be the largest pockets/cavities in a protein (154; 155; 156; 17). We have not examined other cavities within our proteins, but we plan to compare the patterns of valid and invalid ligands in the future. One would assume that the crystal additives on the surfaces of the protein are in shallow pockets with little buried surface area, but covalent cofactors and structural elements of proteins will occupy both surface and buried positions. Patterns of valid vs. invalid ligands of both types should help current efforts in the field to identify binding sites in apo structures. At this time, groups are focusing on the analysis of occupied vs.

Figure 3.7: (A) Distribution of the buried surface area (Å^2) for cavities within Binding MOAD as calculated with GoCAV, note the larger scale for the redundant data. Black bars represent all complexes in Binding MOAD; gray bars represent only the complexes with affinity data. (B) Plots of buried surface area of the cavity (Å^2) vs. ligand size. The data points in black squares are from complexes with Kd data, and gray diamonds are used for complexes with Ki or IC50 data. Error bars for data points were available in two cases. First, if a side chain in the active site was resolved in more than one orientation. Second, some multimer complexes are solved with slight differences in the independent binding sites (for instance, the atomic coordinates of the binding sites within a dimer will not be the exactly same if symmetry was not imposed while fitting the electron density).



unoccupied pockets and having good success (49; 154; 136; 157; 158; 159; 160; 161; 156; 17; 162; 149), but the methods could be further refined with data on the invalid ligands identified within Binding MOAD.

Liang et al. also found that binding sites are either buried cavities or more often pockets with one, occasionally two, exposed openings (17). In agreement with that study, our histograms in Figure 3.8 show that most ligand-binding sites have limited exposure to solvent; GoCAV data shows that 70% of the cavities have $\geq$70% of their MSA buried, and NACCESS data shows that 85% of the ligands have $\geq$70% of their SASA buried. The high degree of burial also parallels findings by Keil et al (161) where they show that binding sites for ligands are deeper and more concave than binding sites for protein-DNA or protein-protein associations. We found that the largest ligands are rarely well buried. They tend to have less percent buried MSA of the cavity and less percent buried SASA of the ligand (Figure 3.9); many of them are short peptide or nucleic acid chains, again fitting with the findings that such binding sites are more shallow.

3.4 Conclusions

The histograms in Figures 3.7A and 3.8 tell us that most ligands are well buried. This fits the common paradigm that many contacts between the ligand and the protein are a significant factor in molecular recognition. Figure 3.9 shows that largest ligands tend to have more exposed surface area. These large ligands are typically peptide, nucleic acid, or sugar chains, and one would expect the patterns of binding such molecules to start to resemble the patterns of proteins binding macromolecules.

The general trends found here do not change with the choice of MSA of the pocket vs. SASA of the ligand. Also, GoCAV and NACCESS use different methodologies and radii, so the patterns appear to be independent of how the calculation is performed. We do want to note that surfaces of the binding site calculated with GoCAV are

Figure 3.8: Histograms of the percent of surface area that is buried. (A) Percentage of buried MSA of the cavity and (B) Percentage of buried SASA of the ligand. Black bars represent all complexes in Binding MOAD; gray bars represent only the complexes with affinity data.

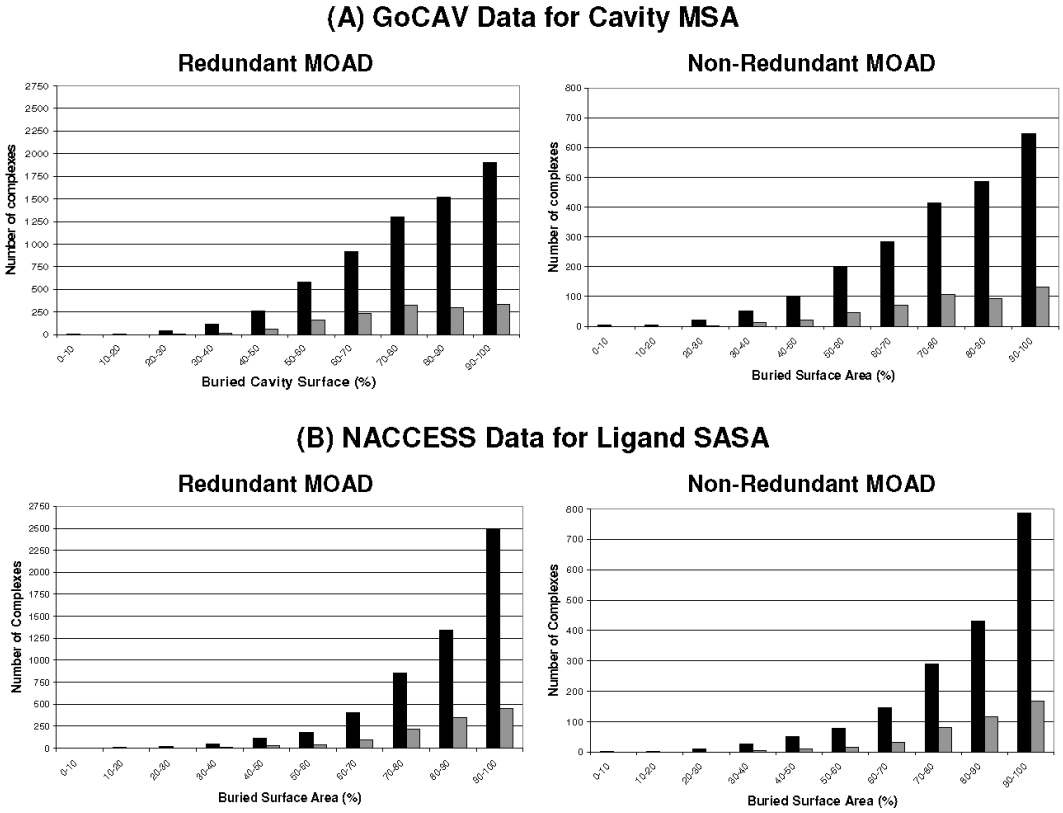
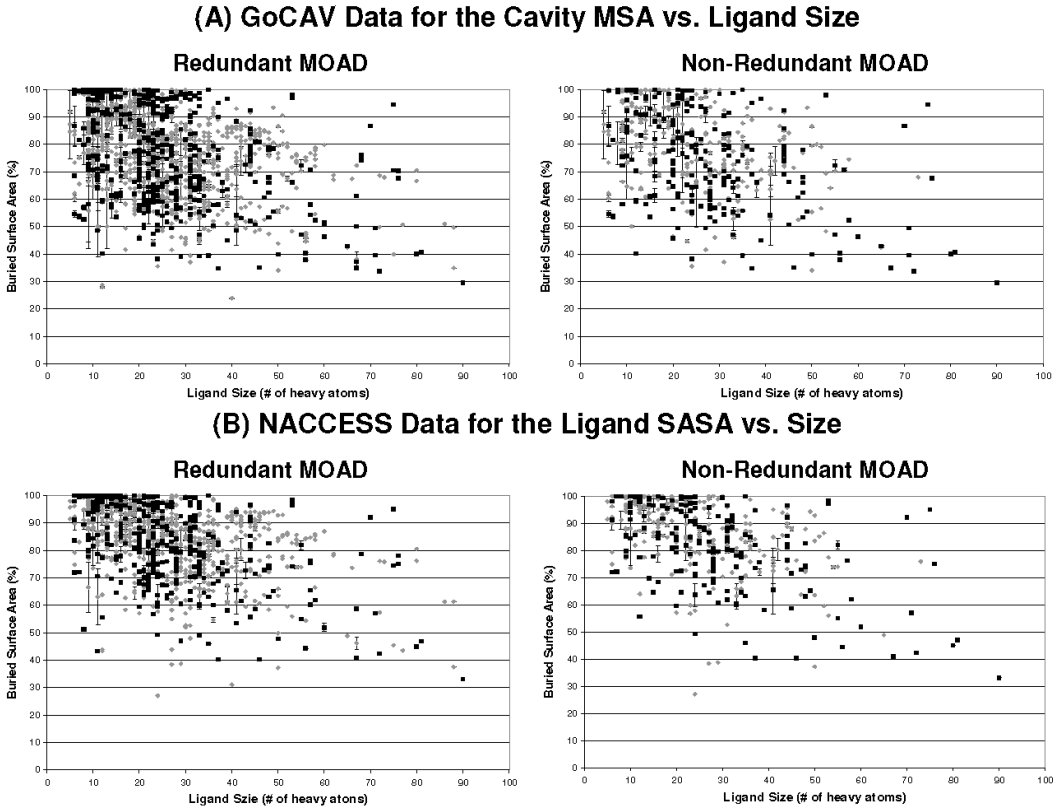


Figure 3.9: The largest ligands tend to have much of their surface area exposed to solvent (low % buried). (A) Percentage of buried MSA of the cavity and (B) Percentage of buried SASA of the ligand. The data points in black squares are from complexes with Kd data, and gray diamonds are used for complexes with Ki or IC50 data.



not completely independent of the ligand because of the use of an ELS to "bound" the binding site. However, the probe never reaches the ELS boundary in a buried binding site. Most of our sites are highly buried, so the majority of the cavity surface is defined only by contacts to the protein. This typically makes the portion of the surface defined by the ELS only a small percentage.

In closing, future efforts with Binding MOAD will allow us to compare – broadly, for the first time – the binding affinity data to the patterns of molecular recognition mined from the PDB. Past studies have mined subsets of the PDB with various structural analyses of proteins and ligands (49; 154; 59; 163; 136; 157; 164; 158; 165; 159; 160; 161; 156; 142; 17; 162; 132; 166; 149; 167), but now, we will be able to add another layer of depth to such studies. There is more to binding affinity than just burying a ligand inside a protein, and all of the complex issues that go into creating an effective scoring function (168) will need to be considered in our analyses. Both shape and chemical complementarity are thought to be the basis of molecular recognition. Our future analyses will have to consider the chemical complementarity or what "types" of surfaces are solvent-exposed or interact with the protein. We will also need to address the very complex issue of entropic changes upon binding.

CHAPTER IV

Differences between high- and low-affinity complexes of enzymes and nonenzymes

4.1 Introduction

Both enzymatic and non-enzymatic proteins can bind small molecules, but enzymes catalyze reactions and have a fundamentally different role from non-enzymes, which may have an impact on their recognition of ligands. Do these two types of binding events have the same physical characteristics? Furthermore, are there any differences between high-affinity complexes and weaker binding events that can be linked to their physical contacts? To answer these questions, physicochemical patterns were mined from our protein-ligand database Binding MOAD (Mother of All Databases), where MOAD is pronounced "mode" as a pun on a ligand's mode of binding.(62; 169)

Binding MOAD is the largest curated database of high-resolution protein-ligand complexes from the Protein Data Bank (PDB).(115) Though it only reflects proteins that can be crystallized, these are the exact systems where structure-based insights will be used. The PDB is the source of all structures used for docking and scoring development by academics. However, the data used here are significantly larger than most sets used to develop existing scoring functions, which are typically sets of <300

complexes of <50 unique proteins. We use 2214 structures: 1790 enzymes and 424 non-enzymes (512 unique enzymes and 176 unique non-enzymes). This study provides an important benchmark of the current landscape available from structural biology (incomplete and/or biased as it may be).

For this study, we have compared distributions of various properties between four classes of protein complexes. Distribution analysis is used widely in many fields, and it is important to stress that it does not define "absolute rules", nor are the data presented as such. These are general guidelines, and of course, there will be exceptions to those trends. Distribution analysis can show that "men are taller than women" and "women live longer than men." Those trends are true even though some women are 6' tall and some men live to 100.

Empirically derived rules can be very useful in discovering and applying new principles in chemistry. One of the most well known examples is Lipinski's Rule of Five, which describes the physical properties of orally-available drugs.(170; 171) These rules provide general guidelines for size, lipophilicity, and hydrogen-bonding characteristics that correlate with the likelihood that a molecule can be orally absorbed into the body. The findings are based on distribution data of the chemical characteristics of orally absorbed molecules going into Phase-II testing. The dataset is biased by issues outside of pharmacokinetics such as the need for good synthesis (not just accessible chemistry, but few steps in high yield) and market considerations (completely economic, no basis in the thermodynamics of protein-ligand binding). The rules do not hold for natural products, actively transported molecules, molecules that require metabolism for activation, or most antibiotics, antifungals, vitamins, and cardiac glycosides. There are plenty of molecules in Lipinski space that are not drugs, and many molecules outside that space that are. Despite these limitations and biases, the Rule of Five is used widely in the pharmaceutical industry.

We hope that the present work will also aid drug discovery. In this study, we

provide new patterns which describe high-affinity, protein-ligand binding and outline differences between enzymes and non-enzymes. Of course, there will be examples that fall outside the typical pattern, but these relationships provide a good description of the general landscape that structural biology can provide at this time. We expect that our understanding will grow as more structures become available through the various protein structure initiatives.(172) These guiding principles may be useful in designing targeted libraries for drug discovery and improving scoring functions. They are also important to advancing our fundamental understanding of chemical biology, protein-ligand binding, and the biophysics that dictate molecular recognition.

Non-covalent, small molecule binding is a tradeoff between the enthalpy gained by making specific contacts between functional groups of the ligand and the protein and entropy lost by forcing the ligand and protein into a specific conformation.(173; 174) Since this study uses crystal structures it is difficult to fully account for the effect caused by entropy. However, it is possible to determine the physical characteristics of the small molecule and the protein which leads to the binding affinity.

Other studies(175; 176) have noted an inherent limitation in mining protein structures for physical characteristics of binding. When a pocket is discovered on a protein surface, it is difficult to identify whether it is a true binding site or if it is capable of high-affinity binding appropriate to represent drug-like binding. This study does not suffer from these limitations; all sites have been curated to assure that they are true binding pockets, and the high-affinity complexes are separated from those with low affinity.

Only complexes with binding data (K_d, K_i, or IC_{50}) were used for this study. No complexes in MOAD are annotated with K_m data, so almost all ligands are inhibitors, agonists, or antagonists (a small number are cofactors, 5%, included only for systems were affinity data is appropriate). We specifically focused on the contacts between the ligand and the protein, excluding any structure with poorly defined contacts such

as missing atoms from under-resolved density or ligands and side chains resolved in multiple orientations. Distributions of ligand size, buried surface area (BSA), exposed surface area (ESA), and other physical characteristics were examined for statistically significant differences between four subsets of the complexes: high-affinity binding to enzymes, high-affinity non-enzymes, low-affinity enzymes, and low-affinity non-enzymes. A common metric to evaluate lead compounds is ligand efficiency.(24; 25; 27; 26) In this study, ligand efficiencies for the different classes of proteins are reported as affinity per size (-ΔG_{bind} divided by the number of non-hydrogen atoms) and per the degree of contact between the ligand and the pocket (-ΔG_{bind}/BSA).

Here, we focus on the most significant differences between molecular recognition of tight and weak binding to enzymes and non-enzymes.

4.2 Methods

Data for this study come from the largest comprehensive database of protein-ligand crystal structures with binding data, Binding MOAD. The latest version of Binding MOAD was created from structures released on 12/31/2008 or earlier; it contains 13138 complexes, comprised of 4078 unique protein families binding 6213 unique ligands. The great care taken in curating this dataset has been outlined elsewhere,(62) but it should be noted for these purposes that ∼11,000 crystallography papers have been examined to determine the appropriateness of every ligand (crystallographic additives, post-translational modifications, and covalently bound ligands are excluded from consideration). From these efforts, binding affinity data is available for 30% of the entries, with a preference for K_d data over K_i data over IC_{50} values. The affinities were converted to free energies of binding by $\Delta G_{bind} = RTln(K_d)$ or simply approximated by $\Delta G_{bind} = RTln(K_i$ or $IC_{50})$ with a temperature of 298 K.

High-affinity binding was defined K_d, K_i, or $IC_{50} \leq$250 nM ($\Delta G_{bind} \leq$-9 kcal/mol),

which is approximately the average of all the complexes with binding data in Binding MOAD. Enzyme complexes were defined from the Enzyme Classification number in the PDB file. The non-enzymes were annotated by hand using keywords reported in the remarks section of the PDB entry. Binding MOAD's high-affinity non-enzymes and enzymes are listed in the Supporting Information. Enzymes that had ligands that were allosteric sites, were considered non-enzymes. For instance, the non-nucleoside inhibitors of HIV Reverse transcriptase are bind in a non-enzymatic allosteric site, and it was included in the non-enzyme list. All complexes and binding data are available at the Binding MOAD website, www.BindingMOAD.org.

To calculate surface areas, BSA and ESA were calculated with GoCAV using radii based on united-atom OPLS parameters.(169) This code reports buried molecular surface area (MSA) of the pocket and also defines ESA of the binding site, bounded by the 3D coordinates of the ligand.

The SlogP for the ligands was calculated using MOE,(177) based on the method developed by Wildman and Crippen.(178) For the 2D and 3D descriptors calculated with MOE, the idealized SDF files from the PDB were used if available; otherwise, the coordinates of the ligand from the protein's structure were taken. Hydrogens were added with MOE. In an effort to identify any differences, all 2D and 3D ligand characteristics available within MOE were compared for the four groups of complexes: high-affinity enzyme, low-affinity enzyme, high-affinity non-enzyme and low-affinity non-enzyme.

4.2.1 Statistical Analysis

Statistical significance was assessed with the programs SAS(179) and JMP(180). Initial assessments used JMP to calculate all pair-wise correlations for the over 200 descriptors calculated. For the descriptors showing interesting trends, the significance of the differences between the distributions of physical properties were determined by

the Wilcoxon rank-sum test, which is most appropriate given the non-Gaussian distributions of the data. We also performed one-way ANOVA, two-way ANOVA, and Tukey-Kramer HSD tests between the four classifications. Since these second series of tests require near-normal distributions, the square-root transform was applied to reduce the skew and bring the distributions closer to normal. For the important descriptors, distribution analyses from JMP are included in the Supporting Information (Supporting Information, Figures A.1-A.7), and each includes the mean, median, quantiles, distribution histogram, and outlier box plot. The results of the Tukey-Kramer HSD test are presented in Supporting Information (Supporting Information Tables A.1-A.5).

Histograms of the distributions of ligand size were binned in increments of 5 heavy atoms. Distributions of BSA and ESA were binned by 50 Å^2. Those plotting ligand efficiency were binned by 0.1 kcal/mol-atom for affinity per size or 10 cal/mol-Å^2 for affinity per degree of contact. Distributions of SlogP were binned by 2 log units. These bin sizes were in proportion to the size of the datasets and were consistent with those automatically generated by JMP.

4.3 Results and Discussion

Considerable effort was made to determine direct mathematical relationships between affinity and surface area, ligand size, or other characteristics of protein-ligand interactions, but there was no global correlation across all complexes. Recent work by Coleman and Sharp(181) based on the PDBbind dataset(57) also found no correlation between affinity and surface area or depth of the binding pocket. Inspired by analyses of distributions of ligand efficiencies from screening data,(24) we changed our approach and focused on distributions of the properties between subsets of protein-ligand complexes.

Table 4.1 outlines the characteristics that differ between high-affinity and low-

affinity binding for enzymes and non-enzymes; all emphasized differences in the datasets have a statistical significance >99.99% ($p<0.0001$) based on a two-tailed, Wilcoxon rank-sum test. Figure 4.1 shows a comparison between each of the subsets of complexes, examining the distribution of ligand sizes, BSA, SlogP, and ESA. Many of the low-affinity complexes have ∼300 Å^2 of BSA, but the high-affinity complexes display more contact. It has been estimated that drug-like binding sites have ∼300 Å^2 of solvent-accessible surface area (SASA).(175) Our measurement for BSA is based on MSA, and so, the slightly higher values of the high-affinity complexes are appropriately comparable.(175)

4.3.1 Different approaches for improving inhibitors of enzymes versus non-enzymes

For enzymes, there is a significant difference in the size of the ligands in high- and low-affinity complexes (Figure 4.1a). High-affinity ligands are much larger (11 more non-hydrogen atoms). However, non-enzymes display very little difference in the size of the ligands between high-affinity and low-affinity complexes (Table 4.1, Figure 4.1b). These differences do not come from any influence of the inclusion of cofactors in the set. The medians are nearly unchanged if they are removed from the dataset (see Supporting Information, Table A.6).

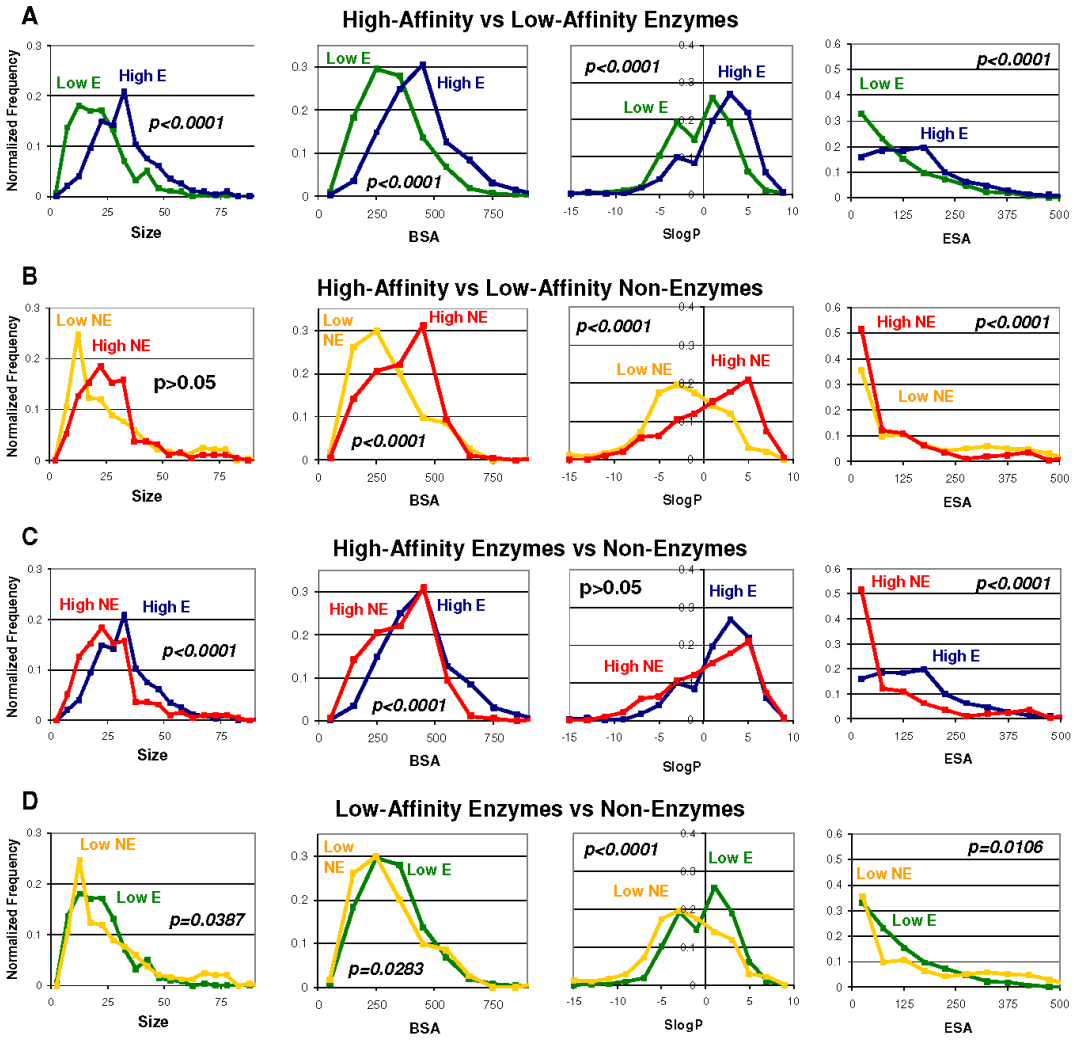
Sizes of the ligands point to a strong difference in the complexes, particularly in how to improve an inhibitor for enzymes versus non-enzymes. To improve the affinity of an enzyme inhibitor, it appears fruitful to add functional groups to increase the complementary contact between the inhibitor and the protein. In contrast, improving ligands for non-enzymes may best involve conservative changes which maintain the ligand's size. Tight binders for non-enzymes are less exposed than the low-affinity ligands, making them more sequestered from the surrounding solvent (Table 4.1). Distributions of the calculated octanol/water partition ratios (Figure 4.1a,b) show

Table 4.1: Characteristics of Protein-Ligand Binding for Enzymes and Non-Enzymes in the Full Dataset.[a]

Median Physical Properties	**Low Affinity** >250 nM ΔG_{bind}>-9 kcal/mol	**High Affinity** ≤250 nM ΔG_{bind}≤-9 kcal/mol	**Comparison**[b]
Enzymes	**1048 complexes**	**742 complexes**	High-affinity ligands are 52% larger and more hydrophobic
ΔG_{bind}	-6.6 kcal/mol	-10.9 kcal/mol	
Size[c]	21 atoms	32 atoms	
BSA	305 Å^2	419 Å^2	
ESA	87 Å^2 (22%)	144 Å^2 (24%)	
(%ESA)[d]	0.3	2.4	
SlogP	0.31 kcal/mol-atom	0.36 kcal/mol-atom	
- ΔGbind/atom	21 cal/mol-Å^2	26 cal/mol-Å^2	
- ΔGbind/BSA			
Non-Enzymes	**234 complexes**	**190 complexes**	Low-affinity ligands are three times more exposed and more hydrophilic
ΔG_{bind}	-7.2 kcal/mol	-10.4 kcal/mol	
Size[c]	22 atoms	25 atoms	
BSA	265 Å^2	361 Å^2	
ESA	118 Å^2 (33%)	45 Å^2 (11%)	
(%ESA)[d]	-2.2	1.5	
SlogP	0.28 kcal/mol-atom	0.41 kcal/mol-atom	
- ΔGbind/atom	22 cal/mol-Å^2	31 cal/mol-Å^2	
- ΔGbind/BSA			
Comparison[b]		Non-enzymes have 17% greater ligand efficiencies	

a. Values presented are medians for each population.
b. All differences noted in the comparisons sections have a statistical significance of >99.99% ($p<0.0001$).
c. Ligand size is given in the number of non-hydrogen atoms.
d. Percent exposure is ESA/(ESA+BSA).

Figure 4.1: Comparisons of (A) enzyme complexes, (B) non-enzyme complexes, (C) high-affinity complexes and (D) low-affinity complexes are presented. High-affinity enzymes are shown in dark blue, and low-affinity enzymes are in green. High-affinity non-enzymes are in red, and low-affinity non-enzymes are in gold. Distribution of ligand sizes (number of non-hydrogen atoms), buried surface area of the pocket (Å^2), SlogP, and exposed surface area (Å^2) are given in normalized percent frequencies. P-values show the significance of the difference in the medians of the distributions, as determined by a two-tailed Wilcoxon rank-sum evaluation (insignificant differences have p>0.05).



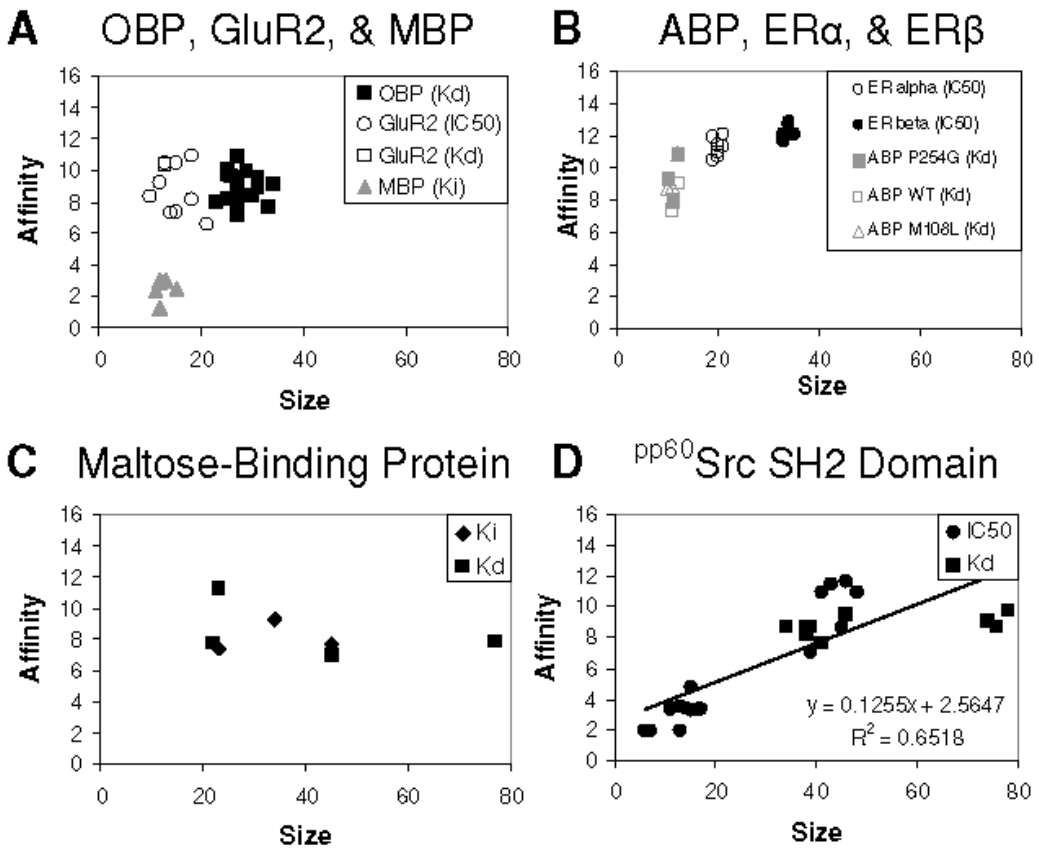
that high-affinity ligands are more hydrophobic than those with low affinity, but there is no significant difference between enzymes and non-enzymes in this regard. It appears that "adding grease" equally improves binding to both enzymes and non-enzymes, consistent with a general desolvation effect.(173)

The above trends for improving inhibitors for enzymes versus non-enzymes come from observing patterns across different proteins (inter-protein relationships), but information to improve inhibitors for a specific target must come from observing trends of one protein binding a variety of ligands (intra-protein binding trends). This is a more difficult comparison to make because few proteins are crystallized with a significant range of bound ligands. For the few that exist, we must divide them into enzymes and non-enzymes, further reducing the sizes of the available datasets. The findings below are qualitative in nature. Overall, our data show that enzymes appear to have better correlations between size and affinity than non-enzymes.

In order to determine a relationship between ligand size and affinity within a protein family (Figures 4.2 and 4.3), the complexes were grouped by 100% sequence identity. This organization ensures that changes in affinity are the result of changes in the ligand and not a mutation within the binding site. (For a few proteins, we were able to combine two sets when the mutations were far from the active site and inconsequential.) Groups that contained $\geq$5 complexes were examined. For non-enzymes, there were only a few proteins available: oligopeptide-binding protein, glutamate receptor 2, estrogen receptor alpha, estrogen receptor beta, arabinose-binding protein, mannose-binding protein, maltose-binding protein, and src SH2-binding domain. For most of the non-enzymes, the ligands are very similar in size and affinity. Six of the eight proteins have a small range of ligand sizes which shows little correlation to affinity (Figure 4.2a, b). The small range of observed ligand sizes supports the idea that conservative changes are most appropriate for trying to improve ligands for non-enzymes. However, the lack of a distinct trend between ligand size and affinity

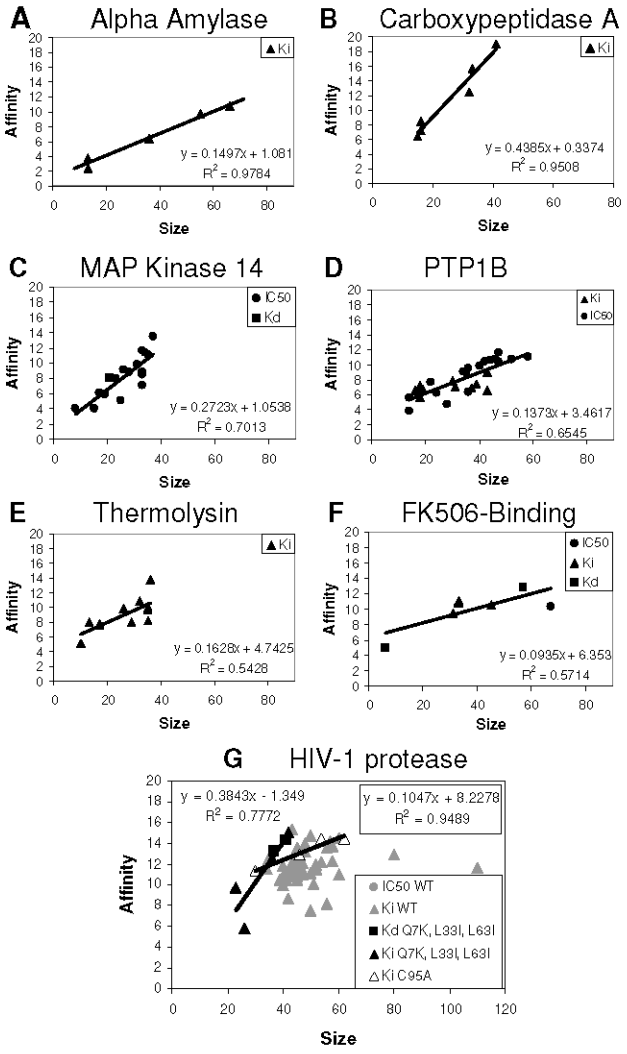
does not necessarily prove that a trend could not be observed. It is unclear if the small range of ligands is the result of the specificity of the protein systems or whether more diverse complexes are simply not available from the PDB.

Figure 4.2: Limited correlation is seen between size and affinity in non-enzymes (A and B). The proteins with "clusters" of points have smaller binding sites and no ligands over 40 non-hydrogen atoms. The ligands have similar sizes and affinities for oligopeptide-binding protein (OBP), glutamate receptor 2 (GluR2) and mannose-binding protein (MBP), arabinose-binding protein (ABP), and estrogen receptor (ER) alpha and beta. The only non-enzymes with a range of ligand sizes are maltose-binding protein and the non-enzymatic site on the SH2 domain of ^{pp60}src tyrosine kinase (C and D, respectively).



Only maltose-binding protein (Figure 4.2c) and the non-enzymatic site on the SH2 domain of ^{pp60}src tyrosine kinase (Figure 4.2d) have a significant range of ligand sizes. The maltose-binding protein complexes contain sugar chains of varying length.

Figure 4.3: Many examples are available of enzyme complexes that show a strong correlation between size and affinity of the ligands; seven are given here (A-G). HIV-1 protease (G) demonstrates that a large collection of ligands may show no correlation, but subsets of data may reveal strong trends (data for the C95A and Q7K/L33I/L63I mutants). It is interesting that even small binding sites with ligands of 40 non-hydrogen atoms or less (B,C,D) show a linear trend with affinity; this was not seen for non-enzymes with small binding sites.



Almost all bind with roughly the same affinity, and this may be explained by the fact that the larger ligands show little difference in the BSA contact, despite the very large range of sizes. The non-enzymatic site on the SH2 domain of ^{pp60}src tyrosine kinase is the only non-enzyme complex showing some correlation between ligand size and binding affinity. It is interesting that the only exception in non-enzymes is a regulatory site on an enzyme. These linear correlations reflect a trend across several ligands, $\Delta(\Delta G_{bind}/\text{size})$, which is slightly different than the ligand efficiency of an individual ligand, $\Delta G_{bind}/\text{size}$. In the discussions below, we will use the term "trend" or "correlation" when comparing across several ligands bound to the same protein, $\Delta(\Delta G_{bind}/\text{size})$.

In the case of enzymes in MOAD, thirty-seven proteins were available with five complexes or more. Unlike non-enzymes, over half of the families showed correlations between size and affinity. For brevity, only seven examples of MOAD's enzymes are given in Figure 4.3. One of the most interesting features of the data in Figure 4.3 is that the slopes - the overall trend for each set - significantly vary! Though a linear correlation can be found for a good number of enzymes, the additive contributions of more functional groups appear to be system dependent, with some contributions being rather small. The trends range from 0.44 kcal/mol-atom for carboxypeptidase A (Figure 3b) to 0.09 kcal/mol-atom for FK506-binding protein (Figure 4.3f). Most scoring functions use additive terms, and these findings underscore the difficulty in developing a universal scoring function, appropriate for all protein systems. Yang *et al.* have also noted these difficulties in development of their M-Score scoring function(182).

However, for 11 enzymes, there was no correlation; the ligands had roughly comparable affinity and sizes, much like the non-enzyme examples. Three enzymes showed a very small range of ligand sizes and a large range in binding affinity (Supporting Information). It is debatable whether these trends are exceptional examples of the

correlation expected for enzymes or whether they indicate cases where only conservative changes in sizes are allowed, as would be expected for non-enzymes. It is also possible that they result from an unusual set of ligands from one chemical class.

Though Babaoglu and Shoichet have used fragments of inhibitors of β-lactamase to show that ligand efficiency is not necessarily additive within a binding site,(183) fragment-based design often couples these small building blocks in the pursuit of high-affinity ligands.(184) From our data above, one might expect greater success for this strategy when targeting enzymes where increasing size generally leads to increasing affinity. A recent study by Hajduk compared fragment-based design for 14 enzymes and four non-enzymes to show that ligand efficiency remained rather constant as the optimal leads were increased in size.(185) The contributions were roughly additive for the best functional groups. The average trend across these systems was 0.3 kcal/mol-atom, with individual systems showing trends from approximately 0.23 to 0.51 kcal/mol-atom (reported as binding efficiency indices of 11-28 pK_d units per MW in kDa). It is encouraging that the values are comparable to the ligand efficiencies reported in Table 4.1.

Hajduk's trends were presented for the most efficient ligands for each protein, emphasizing the most ideal cases of improving a ligand.(185) However, his data for Bcl-xL, a non-enzyme with a large binding cleft, showed that many changes will not be optimal. A detailed analysis for >2300 additional molecules showed that many had significantly lower efficiencies. In fact, he suggests that chemical modifications that reduce the ligand efficiency by >10% deviate too much from the ideal and indicate that either the location or chemical nature of the modification is less desirable.

The HIV-1 protease data (Figure 4.3g) shows that there is a large scatter of inhibitor sizes and affinities, but two subsets of data (from mutants of HIV-1 protease) show strong linearity. This could demonstrate the same issue seen in Hajduk's detailed analysis of Bcl-xL.(185) The full set of data shows wide scatter and little trend, but a

carefully chosen subset could reveal idealized trends for a particular protein system or class of ligands from a specific synthetic series. For HIV-1 protease, the compensation between enthalpy and entropy can be hard to control. Lafont *et al.* have demonstrated that an increase in size from the KNI-10033 inhibitor to the KNI-10075 inhibitor did not increase binding affinity despite a more favorable enthalpy from a strong hydrogen bond.(19) The entropic penalty of changing a thio ether (two heavy atoms) in KNI-10033 to a sulfonyl group KNI-10075 (four heavy atoms) is responsible for the lack of change in binding affinity. That study noted that, although others have been able to optimize certain HIV-1 protease inhibitors with respect to enthalpy, the enthalpy-entropy compensation could make optimization of affinity impossible for some chemical series.

An important caveat should be considered in the preceding discussion. It is possible that strong correlations between size and affinity can only be easily determined for large binding sites. Large ligands can be truncated to provide smaller, weaker ligands that bind to subsites. This would give a wide range of ligand sizes and affinities, allowing a definite size-affinity relationship to emerge from the data. It may be more difficult to determine a trend for a small binding site. This would still imply that enzyme inhibitors are more likely to be improved through the addition of functional groups, simply because the binding sites in enzymes are generally larger than those of non-enzymes. However, if this were the case, the trend would be due to the size of the binding site and not necessarily the protein's basic function.

Though the size argument above is important to note, it is most likely not the cause of the difference between enzymes and non-enzymes. Several examples of smaller binding sites, characterized by ligands of 40 non-hydrogen atoms or less, are presented in Figures 2 and 3. For small non-enzymes, there are no proteins which show a correlation between size and affinity. Conversely, there are several enzymes with small binding sites which do show a good correlation of increased affinity with increased

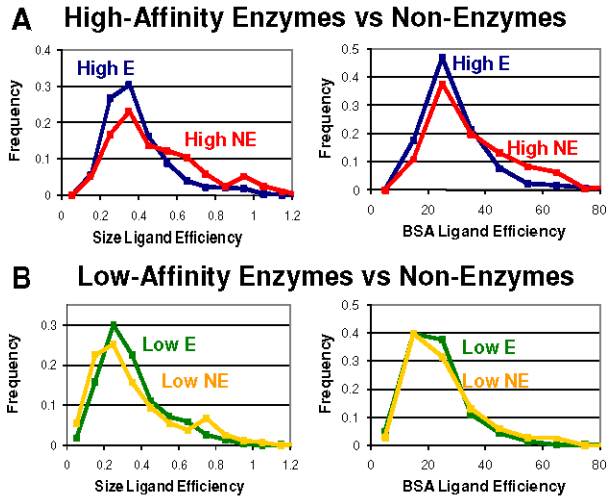
size.

4.3.2 Ligand Efficiencies

Distributions of ligand efficiencies are given in Figure 4.4. Ligand efficiency based on contact (-ΔG_{bind}/BSA) can be compared to established values for the desolvation effect. The free energy of transferring a hydrophobic molecule from a hydrophobic solvent into water has been estimated as 24-47 cal/mol-Å^2, with the higher value being the most widely accepted.(186; 187; 37) Honig and coworkers have noted this is lower than the value of 72 cal/mol-Å^2, derived from the surface tension of a hydrocarbon-water interface.(37) Only 0.8% of the complexes in this study have ligand efficiencies that exceed 72 cal/mol-Å^2 (i.e., greater than Honig's value), and many have efficiencies ranging between 20-40 cal/mol-Å^2. The low-affinity complexes are roughly bounded by the 47 cal/mol-Å^2 value (only 4.1% have greater efficiencies), but the high-affinity complexes have large populations greater than that value. Although, the complexes in Binding MOAD are not exclusively driven by hydrophobic association, these values provide a yardstick for comparisons. However, it should be noted that the range of values from the literature are based on SASA of small molecules in differing environments (ligands), and our values are based on MSA of the contacts within the pockets. While the comparison is not ideal, MSA-based values for ligands are not prevalent in the literature, and SASA of a pocket is not equivalent to SASA of a ligand.

For low-affinity complexes, the ligand efficiencies are basically the same for enzymes and non-enzymes (Table 4.1, Figure 4.4b). However, the differences are significant in high-affinity complexes ($p < 0.0001$ for both efficiencies). The ligand efficiencies for high-affinity, non-enzyme complexes are $\sim$17% greater than those of high-affinity, enzyme complexes (Table 4.1). Non-enzymes in Figure 4.4a show a broader distribution of efficiencies and much higher populations above 0.4 kcal/mol-

Figure 4.4: Distribution of ligand efficiencies per size (-kcal/mol-atom) and per contact (-kcal/mol-Å^2), given in normalized percent frequencies. Distributions present comparisons of (A) high-affinity complexes ($p<0.0001$ in both cases) and (B) low-affinity complexes. High-affinity enzymes are shown in dark blue, and low-affinity enzymes are in green. High-affinity non-enzymes are in red, and low-affinity non-enzymes are in gold.



atom (55% of high-affinity non-enzyme complexes vs 37% of high-affinity enzyme complexes) and 30 cal/mol-Å^2 (51% of non-enzymes vs 35% of enzymes). *On average over the high-affinity complexes, every atom and square Ångstrom of buried cavity surface is worth more free energy in non-enzymes!*

The differences in efficiencies between high-affinity enzymes and non-enzymes are not dependent on the choice of cutoff between high- and low-affinity complexes. Even if the full set of enzymes is compared to the full set of non-enzymes, the ligand efficiencies are better for non-enzyme complexes. For the 1790 enzyme complexes, the median ligand efficiencies are 0.33 kcal/mol-atom and 23 cal/mol-Å^2; the median ligand efficiencies for the 424 non-enzymes are 0.36 kcal/mol-atom and 26 cal/mol-Å^2.

The same patterns for enzymes and non-enzymes are observed when redundancy is removed (Appendix A, Table A.7, Figures A.8 and A.9). This is important because it corrects for some biases in the dataset by using only one complex of a protein (some proteins have hundreds of entries and are heavily represented in the PDB). The

non-redundant dataset in Binding MOAD is obtained by grouping the proteins into families of 90% sequence identity and representing that family by the single complex with the highest-affinity ligand - in essence, the optimal binding event available for that individual protein. There are 688 unique complexes in this dataset, 512 enzymes and 176 non-enzymes. Again, the high-affinity enzymes (235 complexes) have poorer ligand efficiency than the high-affinity non-enzymes (85 complexes). For the non-redundant datasets, the median ligand efficiencies for high-affinity enzyme complexes are 0.39 kcal/mol-atom and 28 cal/mol-Å^2. The median ligand efficiencies for the non-redundant, high-affinity, non-enzyme complexes are still larger at 0.44 kcal/mol-atom and 34 cal/mol-Å^2. The smaller number of complexes produces nearly identical distributions, and although the p-value of the comparison is slightly poorer ($p = 0.04$), it is still significant (96%).

4.3.3 Efficiencies, evolution, and druggability

The significant differences in ligand efficiencies suggest a differentiation in the binding sites of these two classes of proteins, based on their function. This may reflect the different evolutionary pressures upon enzymes and non-enzymes. The higher ligand efficiencies of non-enzymes make them, in essence, more responsive to low concentrations of ligand molecules. This is fitting, given their roles in signaling and regulatory control of cellular function in response to stimuli. Conversely, enzymes are optimized to bind molecules, change them, and release them again.

Ligand efficiencies are one key factor in describing the druggability of a target. Does this imply that non-enzymes may be more druggable? In general, higher ligand efficiencies mean that drug-like affinities can be obtained with smaller molecules. Smaller molecules would tend to provide better oral absorption and fewer functional groups for toxicity concerns.(188; 176; 189; 190) Of course, ligand efficiencies reflect "bindability", and it is important to recognize that there are additional properties

that make a protein a suitable drug target. It must be essential to the disease state. Leads must show selectivity to avoid any negative consequences of off-target binding events. There are a myriad of ADME and pharmacokinetic properties to be considered. However, the differences in ligand efficiencies do indicate a greater likelihood to have better drug-like properties for inhibitors, agonist, and antagonists of non-enzyme targets.

Many non-enzymes are the subject of intense drug discovery efforts in both the private and public sectors; for instance, hormone receptors, signaling proteins, and transcription regulators are targets for anticancer treatment.(191; 192) Recent discussions on the druggability of protein-protein interfaces note that these difficult targets may be more amenable than originally thought.(193; 111) Small molecules have been developed that bind to key hot-spot regions with greater efficiencies and deeper burial than the natural partner. Furthermore, many of the non-enzymes not represented in the PDB are membrane-bound receptors. Even though they are not included here, it is likely that the additional information would support the hypothesis that non-enzymes are more druggable, since they are the target of many drugs. G-protein coupled receptors alone constitute 30% of the drugs on the market,(189) and genomic analysis has indicated many more receptors are druggable.(194)

Our results are also in good agreement with a recent study that estimated the druggability of 1096 non-redundant human proteins.(176) The predictions used a statistical model trained on NMR-screening data using a small fragment library.(195) Four of the top six classes were non-enzymes: vitamin-binding, steroid-binding, lipid-binding, and nucleotide-binding proteins.(176) The non-enzymes that were predicted to be the least druggable were large macromolecular complexes and are not reflected in Binding MOAD and this study.

4.3.4 What produces the higher ligand efficiencies in non-enzymes?

Obviously, the root cause of the disparity in ligand efficiencies between enzymes and non-enzymes is of paramount interest. Though the ligands for non-enzymes are smaller, the SlogP characteristics are roughly the same for high-affinity ligands of enzymes and non-enzymes (Figure 4.1c). *If the ligands are chemically similar, then the difference in efficiencies must come from the protein pocket.* The most significant difference is the degree of exposure for ligands of non-enzymes versus enzymes. High-affinity ligands have a median exposure of only 11% in non-enzymes, but 25% in enzymes (note that %ESA are used instead of ESA to correct for the difference in sizes of the ligands). Low-affinity ligands for non-enzymes are significantly more exposed (median of 33%), even more than the low-affinity ligands for enzymes (22%). Tight and weak inhibitors have the same degree of exposure in enzymes, but tight ligands for non-enzymes are much more encapsulated than the weak ligands ($p<0.0001$). Other 2D and 3D ligand descriptors displayed no significant patterns. This comparison was cognizant of correlations between characteristics; for instance, differences in surface area are correlated to size and were not "double counted" as additional differences between high-affinity ligands of enzymes vs non-enzymes.

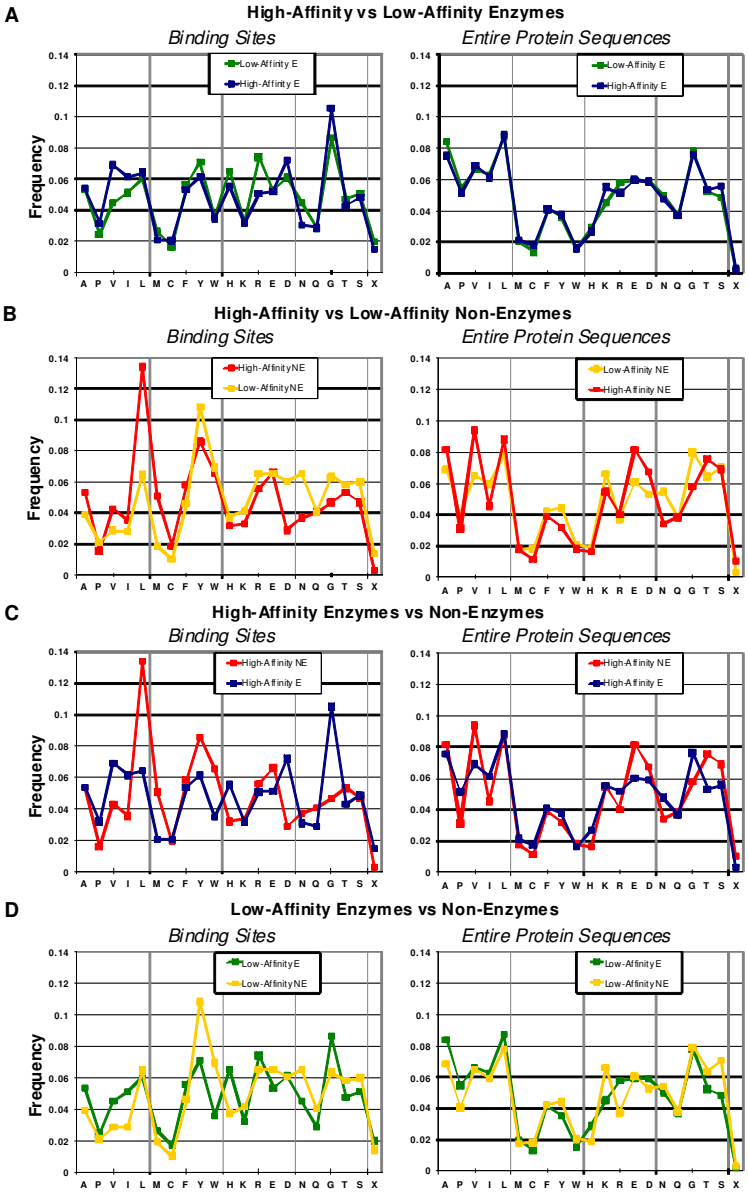
Amino acid composition of the binding sites was examined (Figure 4.5, left column). There is little difference between the binding sites of high- and low-affinity enzyme complexes. The largest differences are an increase in Val content in high-affinity enzymes and an increase in Arg in the low-affinity complexes. For enzymes, the hydrophobic residues (Ala through Trp) on Figure 4.5 are 47.0% of the binding sites for high-affinity complexes, but 43.9% for low-affinity ones. This is fitting with the aforementioned finding that the high-affinity ligands are slightly more hydrophobic. The comparison between binding sites of high- and low-affinity non-enzyme complexes shows more pronounced variation, but also holds the general pattern of high-affinity complexes having more hydrophobic content. The Ala-Trp residues are 55.9% of the

binding sites for high-affinity complexes, but 43.2% for low-affinity ones. What is most interesting is the comparison between enzymes and non-enzymes, particularly for the high-affinity complexes. The hydrophobic content is higher for non-enzymes (55.9% vs 47.0%), but the reader should recall that there is no significant difference in the SlogP of the ligands (in fact, the median value for non-enzymes is more hydrophilic). Why are more hydrophobic sites recognizing slightly more hydrophilic molecules with better affinity? The answer may lie in the fact that the amino acids making the contacts are significantly different. In high-affinity non-enzymes, Leu and Met provide a large portion of the hydrophobic contacts, at the expense of Val and Ile. The non-enzyme's preference for Glu over Asp is reversed in high-affinity enzyme complexes, yet the use of Lys and Arg is the same. Leu, Met, and Glu are larger than their counterparts Val, Ile, and Asp. It is possible that those residues are slightly more polarizable. (Confirmation will have to come from in-depth examinations of fully modeled complexes, inclusive of added hydrogens, detailed atom typing, and possibly polarizable force fields. To do this for thousands of complexes is a sizable effort, and outside the scope of the present study.) It should be noted that differences in the binding sites are not correlated with differences in the overall amino acid content; the reader should compare the left and right columns in Figure 4.5. Leu, Met, Phe, Tyr, and Trp make up nearly the same percentage of residues in the protein sequences, but not the binding sites. This selective placement of differing residues within binding pockets may have direct relevance to analyses of hot-spot regions and potential binding sites on proteins.(196; 197; 198)

4.3.5 Most druggable enzymes

Of course, many pharmaceutically relevant targets are enzymes. By no means is it suggested that they are not appropriate drug targets, especially when they constitute 47% of the drugs on the market(189) and a large percentage of new targets identified

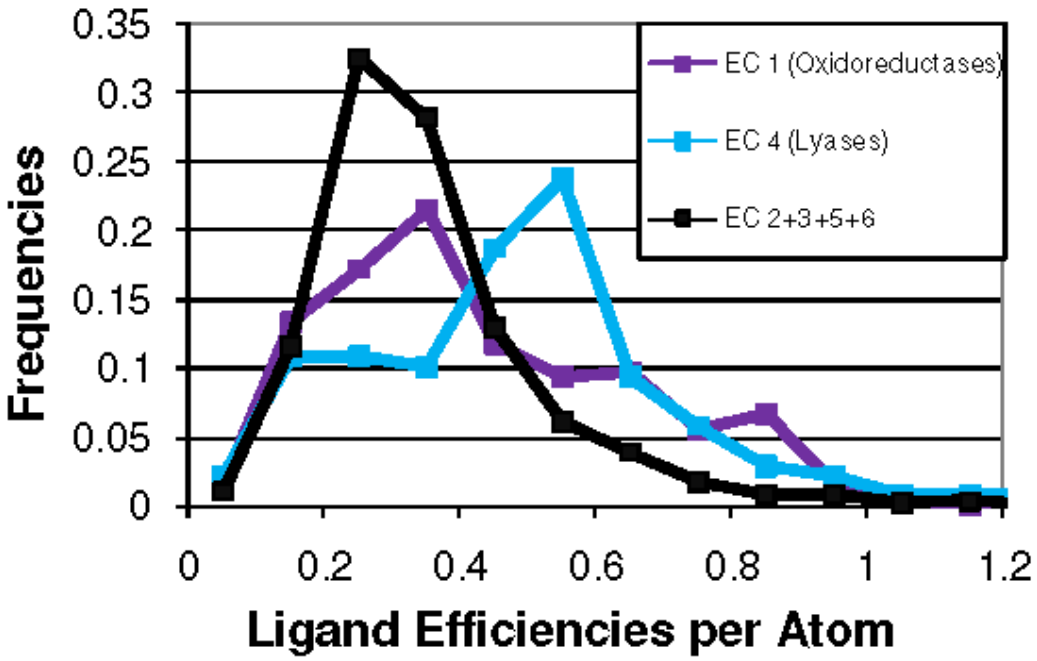
Figure 4.5: The binding sites (left) and the entire protein sequences (right) are analyzed for amino acid content. Distributions are given in normalized frequencies percent frequencies. Amino acids within 4Åof the ligands are considered to comprise the binding site. Distributions of (A and B) low- and high-affinity complexes of the same class show smaller differences than comparisons between enzymes and non-enzymes (C and D). Amino acids are listed by hydrophobic, aromatic, cationic, anionic, and hydrophilic nature. "X" denotes contacts with cofactors, unnatural amino acids, and covalent modifications on the protein.



through genomic analysis.(194) The distribution of ligand efficiencies for the enzyme classes suggests that lyases and oxidoreductases are the most druggable enzymes, Figure 4.6. The distribution of lyases is significantly shifted to higher efficiencies, standing out from the other data. The better efficiencies for oxidoreductases come from an increased population in the tail of the distribution. The median ligand efficiencies for the 139 lyases are 0.50 kcal/mol-atom and 33 cal/mol-Å^2; and the median ligand efficiencies for the 256 oxidoreductases are 0.39 kcal/mol-atom and 26 cal/mol-Å^2. The 1395 enzymes from the other four classes have median efficiencies of 0.31 kcal/mol-atom and 23 cal/mol-Å^2, which are significantly lower (significance of $\geq$99.99% using the Wilcoxon test). It should be noted that the two enzymes which were predicted to be most druggable in the aforementioned study were also lyases and oxidoreductases, in that order.(176)

Recently, a new method was introduced to predict druggability of a binding site by estimating the site's maximum K_d based on the percent hydrophobic SASA and a scaling factor for efficiency that is dependent on the curvature of the site.(175) The model was trained on 8 enzymes and applied to 63 structures, comprised of complexes of 26 enzymes and a single structure of the non-enzyme mdm2.(199) An important goal of the study was to fit a predictive equation to assess druggability of a site based on protein-ligand structures of orally available compounds. This feature of the study is important to note because the contributions of various physical characteristics within the model should reflect both high-affinity binding and oral bioavailability of the ligand. The model was fit under the assumption that hydrophobic desolvation is the major driving force of binding, so terms based on electrostatics were not included. The model was able to properly rank the training set, noting that outliers were compounds with strong electrostatic components, prodrugs, or ligands that are actively transported. The model was then used to identify new, druggable structures from the PDB. It was interesting that the two newly identified targets were both

Figure 4.6: Distribution of ligand efficiencies (-kcal/mol-atom) for enzymes, given in percent frequencies normalized for the different number of complexes in each enzyme class. The distribution of transferases (EC 2, 468 complexes), hydrolases (EC 3, 843 complexes), isomerases (EC 5, 60 complexes), and ligase (EC 6, 17 complexes) are the same and have been added together for this example (black line). Oxidoreductases (EC 1, purple line, 256 complexes) have larger populations in the higher efficiencies ($p<0.0001$). The distribution of lyases (EC 4, blue line, 139 complexes) is notably shifted ($p<0.0001$).



enzymes. With only two new targets presented, it is not clear whether the model preferentially identifies enzymes over non-enzymes, but a preference towards enzymes may be expected from their model given the training and test sets used. Our data indicate that enzymes and non-enzymes may require different models in such analyses. Furthermore, many of the ligand efficiencies in our set exceed the established values for hydrophobic association, indicating that the most efficient complexes have additional factors which contribute to their affinity. The affinity of these complexes may not be well described by models based solely on hydrophobic SASA.

4.4 Conclusion

We have presented a substantial mining study of Binding MOAD, the largest public database of curated protein-ligand structures with binding data. Physical characteristics of bound ligands were compared between enzymes and non-enzymes as well as high-affinity and low-affinity complexes. The comparison between ligand sizes for low-affinity versus high-affinity binding shows that divergent approaches are likely needed to improve the affinity of enzyme inhibitors versus those for non-enzymes. The traditional approach of adding functional groups to fill more of the pocket may work for enzymes, but it may not be as appropriate for non-enzyme systems. However, making ligands more hydrophobic appears to aid binding in both enzymes and non-enzymes.

Non-enzymes have higher ligand efficiencies than enzymes, which may be a reflection of their biological roles. This is also encouraging when considering the druggability of non-enzymes. In the pharmaceutical industry, ligand efficiencies have become a metric for evaluating hits from screening campaigns and even candidate compounds.(25) Our results would caution against applying a rigid standard across all protein targets. At the very least, a cutoff based on ligand efficiency should differ between enzymes and non-enzymes. Ideally, cutoffs would differ between protein

families and only be considered as one of several guidelines in a selection process.

Binding MOAD provides strong support of several mathematical models cited above,(199; 185; 176) particularly those of Hajduk and coworkers. Our results have implications for the development of scoring functions for docking and predicting druggability of a binding site.(200; 201; 202; 203) The differences between non-enzymes and enzymes, as well as the differences across enzymatic systems, underscore the challenges of developing universal functions that perform well across all systems. Modest improvement might be achieved by developing separate functions for enzymes and non-enzymes, with even greater improvement expected for functions trained on specific protein families.

CHAPTER V

Charge-charge interactions appear to dictate the maximum ligand efficiencies available for protein-ligand binding

5.1 Introduction

Protein-ligand binding is a delicate balance between the loss of entropy resulting from complexation and the enthalpy gained by forming favorable contacts with the protein (4; 19). The precise contribution of these contacts is a source of debate and has provided a significant obstacle in the ability to predict how small molecules will bind (204; 182; 205). The interplay between entropy and enthalpy is difficult to determine since they are influenced by several factors. For entropy, binding two entities results in a loss of six degrees of freedom, a change in the internal flexibility of the protein and ligand must be taken into account, and the reorganization of water around the ligand and within the binding site has significant implications. In the case of enthalpy, several types of contacts can be made to varying degrees in the binding site (4). Current thinking is that van der Waals forces are the most significant factor for binding due to tight packing between the small molecule and protein (16; 4). Hydrogen-bonding and electrostatic interactions are thought to contribute more to the specificity of binding (4). Since these interactions are also present with water and

counter ions, they are thought to have a smaller impact on affinity (4).

Highlighting the different interpretations regarding the drive for efficient binding, there has been contradictory evidence as to which types of interactions play the most significant roles in the binding of biotin to streptavidin, the tightest known natural complex. In 1993, Miyamoto and Kollman used free energy perturbation on biotin•streptavidin and N-L-acetyltryptophanamide•α-chymotrypsin to show that the increased binding affinity for the biotin-streptavidin system can be accounted for by van der Waals contacts made in the biotin•streptavidin complex where the pocket in streptavidin is preformed as in the traditional lock-and-key theory (21). However, newer work has shown that networks of hydrogen bonds are responsible for the strong binding in the biotin•streptavidin complex (22).

A common metric to evaluate a small molecule's ability to bind is "ligand efficiency". This metric is defined as binding affinity per number of non-hydrogen atoms (24; 25; 26). It was first introduced by Kuntz *et al.* in 1999 (27), where they analyzed 159 tightest-binding complexes and the relationship between the number of heavy (non-hydrogen) atoms present in a ligand and its affinity. They showed that each heavy atom can provide at most -1.5 kcal/mol of binding affinity (27). This maximum was consistent with their predictions of the maximum affinity obtainable by van der Waals and hydrophobic interactions (27). Though many of the most efficient ligands were metals and small ions, electrostatics was given little attention. Even in recent investigations this class has been ignored because they are not "drug-like" and most scientists prefer to focus on drug-like molecules for ligand efficiency (206; 207; 208).

In this study, we investigated which properties lead to an optimal efficiency. To study general patterns with regard to binding affinity and efficiency, it is necessary to use a large set of protein-ligand complexes for which a structure has been solved and an experimentally-derived binding constant (K_d, K_i, or IC_{50}) has been determined.

We used the largest dataset available, Binding MOAD(62; 169), to explore the relationship between structure and binding affinity, extending Kuntz's examination to include all available binding events in the Protein Data Bank (115). By looking at the most efficient ligands and the characteristics of their binding pockets, we reveal which interactions are most important to provide the highest binding affinity and efficiency. This study explores all binding events with the goal of examining fundamental biophysical properties, rather than focusing solely on properties of drug-like chemical space.

5.2 Methods

Structural properties were derived from the complexes in our protein-ligand database Binding MOAD (Mother of All Databases) (62; 169). Binding MOAD is the largest database of high-resolution protein-ligand complexes annotated with binding data from the PDB(115) (13,138 complexes comprised of 4078 unique protein families, binding 6213 unique ligands). We have compiled binding affinity data for 32% of the entries (4203 complexes), with a preference for K_d data over K_i data over IC_{50} data. The free energy of binding was determined directly from K_d values by ΔG_{bind} = -RT$\times$ln(K_d), and in the case K_d was not available, we approximated the free energy of binding using ΔG_{bind} = -RT$\times$ln(Ki or IC_{50}). All structures and affinity data are freely available at http://www.BindingMOAD.org.

Only complexes with binding data were used for this study. Coordinates were taken from the biological unit files provided by the PDB, which display the functional form of the protein. These files were processed to remove artifacts. We specifically focused on the size of the ligand and its contact surface with the protein, so any structure with poorly defined contacts were not considered. Therefore, we excluded structures with partially occupied or missing atoms from under-resolved ligands or side chains, as well as structures with too many atoms from ligands or side chains

resolved in multiple orientations. A ligand was determined to have too many or two few atoms if the number of atoms in the formula did not match the number of atoms in the coordinate section of the pdb file. The total number of structures used in this study was 2794.

Ligand efficiency is the free energy of binding divided by the number of non-hydrogen atoms in the ligand (24; 25; 27; 26). Hence, a ligand with 10 atoms is twice as efficient as a ligand with 20 atoms if they bind with the same affinity. In this study, ligand efficiencies are reported as affinity per size (-ΔG_{bind}/atoms) and per degree of contact between the ligand and the pocket (-ΔG_{bind}/BSA).

Surface areas were calculated using OPLS-based radii(150) with our code GoCAV which reports buried molecular surface area (BSA) of the pocket (169). Variation in BSA occurs when several examples of ligand binding occur in the biological unit (i.e., slightly different interactions for three ligands in the three binding sites of a homotrimer). This variation is represented by error bars on the graph of BSA. The exposed surface area (ESA) is also computed from the total surface area minus the BSA.

To estimate the electrostatic interactions of the ligand and the pocket with respect to efficiency, we calculated the minimum distance of each charge of the ligand to the charged residues of the pocket, including any metal atoms that may be present in the binding site. We then averaged the minimum distance over all the charge sites on the ligand. The charge sites were determined by calculating the pKa of each atom in Pipeline Pilot(209) with the pKa calculator at a pH of 7.0.

5.3 Results and Discussion

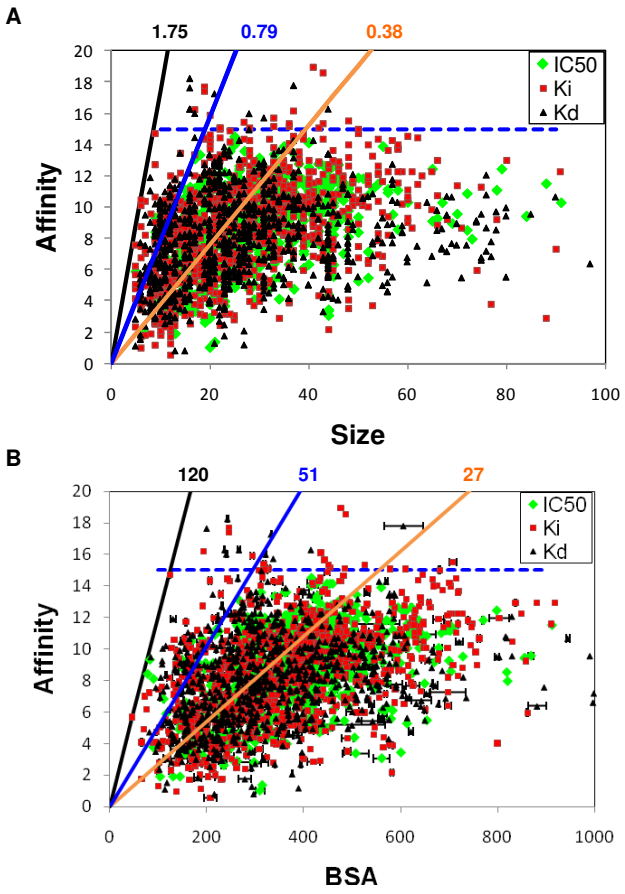
5.3.1 Maximum and average ligand efficiencies

If van der Waals terms are the definitive contribution, then we may expect to see a correlation between affinity and contact surface area between the protein and ligand. However, no correlation is seen between affinity and size or conatcet area(Figure 5.1A & 5.1B).

Our dataset is significantly larger than that of Kuntz *et al.* (27), and we find a slightly higher maximal efficiency for ligands of -1.75 kcal/mol-heavy atom. This "hard" limit is set by several systems, but an alternative "soft" limit is -0.79 kcal/mol-heavy atom, which is the upper bound of 95% of the data in Figure 5.1. The soft limit is established by a significant number of K_d measurements. Given the significant drop in efficiency between the two limits, it is extremely rare to find exceptional ligands and suggests that the -0.79 kcal/mol is a sufficient limit for most uses.

The average and median efficiencies of our dataset are -0.38 kcal/mol-atom and -0.33 kcal/mol-heavy atom, respectively. These averages are in agreement with average values for ligand efficiency of -0.37 kcal/mol-heavy atom for enzymes (median = 0.33 kcal/mol-atom) and -0.42 kcal/mol-heavy atom for non-enzymes (median = -0.36 kcal/mol-heavy atom), as reported in our previous work (210). Accurate benchmarks for ligand efficiencies are very important because these values define physical limits of ligand binding. Furthermore, ligand efficiencies are often used to evaluate HTS data or to eliminate lead compounds during a drug development cycle (24; 25; 26; 211). Anecdotally, the best ligand efficiencies from HTS data approach -0.6 kcal/mol-atom (26; 211). Pushing for leads with ligand efficiencies near -0.3 or -0.4 kcal/mol-atom from a simple combinatorial library may be too restrictive for some systems as this is near the average for all good structures, as noted above (25). However, ligand efficiencies of candidate compounds must often be higher to allow for changes during

Figure 5.1: Plotting the affinity of the complexes versus their physical characteristics reveals the limiting cases as well as the general trends. Measurements used for affinity data are noted as IC_{50} (green diamonds), K_i (red squares), or K_d (black diamonds). (A) Affinity versus size of the ligand. Affinity is given in -kcal/mol and size is given as the number of non-hydrogen atoms. The units of the ligand efficiencies listed above the lines are -kcal/mol-atom. (B) Affinity versus the buried surface area of the binding site, in $Å^2$. The units of the ligand efficiencies listed above the lines are -cal/mol-$Å^2$. The "hard limits" of ligand efficiency are denoted with black lines and values; the "soft limits" which bound 95% of the data are denoted with solid blue lines and values; the average ligand efficiencies are given with orange lines and values. The dashed blue line denotes how few of the complexes have affinities greater than 15 kcal/mol.



further drug development (211; 212).

We can also define ligand efficiency in terms of BSA of the binding site. Others have proposed metrics for ligand efficiency based on free energy of binding per surface area of the ligand, but these have been based on pharmacokinetic considerations and are not equivalent to contact surface area between the ligand and its protein target (24; 25; 26). Recently Nissink, has proposed that the maximal ligand efficiency should be proportional to protein-ligand contact area and volume (206). That work further suggests a modified measure of ligand efficiency based on affinity/$N^{3.0}$, to estimate the area to volume ration of a ligand (206). This metric is also useful for reducing the dependency of a traditional ligand efficiency based on affinity/N, where N is the number of heavy atoms (206).

Estimates based only on the ligand ignore a large portion of the interaction with the protein. Instead, we have chosen to measure the contacts directly. In our description based on the BSA of the binding site, the average efficiency is 27 cal/mol-Å^2. Houk and coworkers coupled structure and affinity data for a moderate set of over 1000 host-guest, 175 antibody-antigen, and 176 enzyme-inhibitor complexes to propose that affinity is proportional to BSA of the ligand (213; 214). Their data implies a relationship, equivalent to 7 cal/mol-Å^2 (reported as approximately 1 log K_a for every 90 Å^2 of buried surface). This average is approximately one-fourth of our average, but Houk's trend is for surface area of the ligand and ours is for molecular BSA of the binding site. Other reported values of the relationship of surface area versus free energy for transferring a hydrophobic solvent into water range from 24 to 47 cal/mol-Å^2 (186; 187), which is in excellent agreement with the range between our average and soft-limit efficiencies.

In Figure 5.1B, the "hard limit" for efficiency is 120 cal/mol-$\mathring{A}^2$ and the soft limit that bounds 95% of the data is 51 cal/mol-$\mathring{A}^2$. We were surprised to find that the maximum efficiency with respect to BSA was in exact agreement with limits

proposed for macromolecular binding (215). In a follow-up work examining protein-protein, protein-RNA, and protein-DNA complexes, Brooijmans *et al.* established the same limit of 120 cal/mol for every $\mathring{A}^2$ of BSA (215). Macromolecular recognition generally involves large, flat regions of a protein surface (216), but despite that large contact surface, macromolecules do not inherently bind with higher affinities than small molecule ligands (215). Keil *et al.* have shown that binding sites for ligands are deeper and more concave than binding sites for protein-DNA or protein-protein associations, implying a good degree of burial for small molecules despite their smaller size (161). It is rather remarkable that the 120 cal/mol-$\mathring{A}^2$ limit of binding efficiency appears to be universal across all varieties of binding interfaces on proteins.

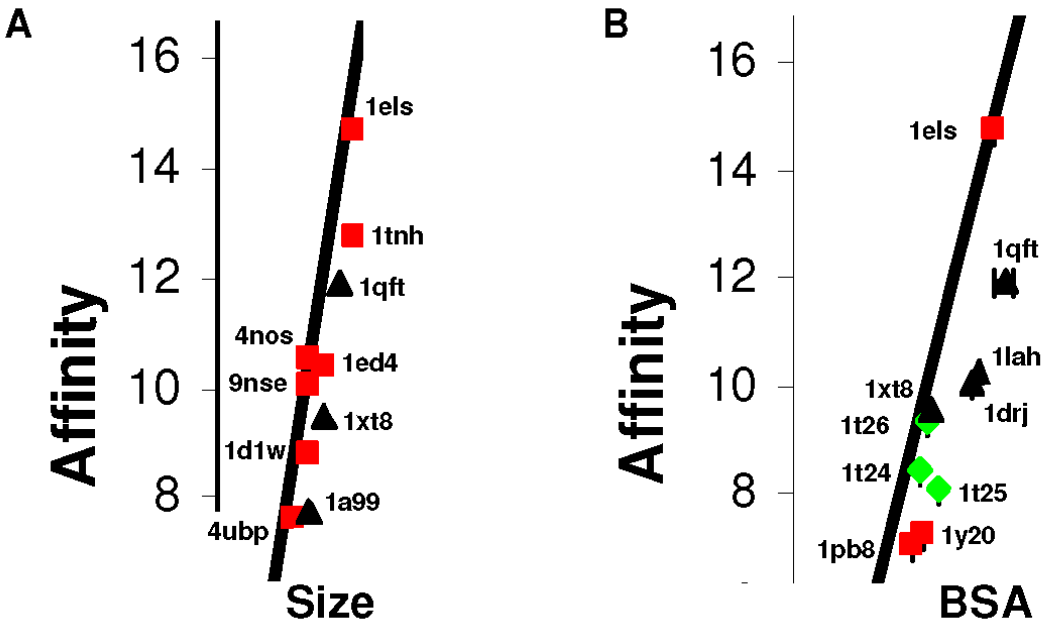
5.3.2 Electrostatic Interactions Define Maximal Efficiency

Structures which define the limit of ligand efficiency all share a single distinct characteristic: every system involves a charged ligand, in contact with a charged protein residue or a metal ion cofactor. In fact, many of the ligands with the best efficiencies have two or three charge centers, and they are complemented in their binding sites by several charged side chains and/or dicationic ions. Figure 5.2 shows the systems with the maximum efficiencies, annotated with their PDB codes. The highest efficiency is seen for a phosphonoacetohydroxamate compound with a -3 charge that is sandwiched between two dications in yeast enolase (PDB code 1els Figure 5.3) (217). The crystal structure shows several unusually tight contacts in the chelation (2.1 Å) which create very small contact surfaces. Not only is the small molecule bound by two magnesium ions, there are two charged aspartates, two glutamates, two lysines, an arginine, and a histidine (that potentially could be charged) in the vicinity.

Other high efficiency complexes include a charged benzylamine coordinated to an acidic side chain in trypsin (1tnh) (218), a dicationic histamine complexed by four acidic side chains in tick histamine-binding protein (1qft) (219), nitric oxide

synthase binding +1-charged isothioureas (4nos, 1ed4, 1d1w, and 9nse - the natural substrate for this enzyme is arginine, which has a positively-charged side chain and a zwitterionic core) (220; 221), a zwitterionic cystine complexed by four charged side chains in the cystine transporter (1xt8) (222), a +2-charged 1,4-diaminobutane in the putrescine receptor (1a99) (223), and an anionic acetohydroxamic acid inhibitor sandwiched between two Ni^{+2} in urease (4ubp) (224). Each of these binding sites can be viewed in Figure 5.3. Even though some of these structures contain metal ions and may be considered partially covalent by some, each structure in Binding MOAD has been verified to be non-covalently bound, according to the primary citation listed in the PDB for the structure (62).

Figure 5.2: Close up view of the complexes with the highest ligand efficiencies. (A) Affinity (kcal/mol) compared to size as in Figure 5.1A. (B) Affinity compared to BSA as in Figure 5.1B. Complexes are labeled with their PDB codes.



Examining the structures that are at the maximum limit of efficiency per BSA shows three structures in common with efficiency per non-hydrogen atom (1els, 1qft, and 1xt8). We note that all but one of the additional systems in Figure 5.2B contain

Figure 5.3: Binding sites of the 11 most efficient complexes. Figures show all residues within 4Å of the small molecule ligand. The ligand is colored by atom type. The water is colored red and shown in small spheres. Metal ions are shown in larger blue spheres. Acidic residues (Asp, and Glu) are colored red; basic residues (His, Lys, Arg) are colored blue; hydrophobic residues (Ala, Ile, Leu, Met, Phe, Pro, Val) are colored green; hydrophilic residues (Cys, Gly, Asn, Gln, Ser, Thr) are colored white; and Tyr and Trp are colored either green or white depending on the interaction made with the ligand. The heme is colored with C=light blue and the Iron=brown.

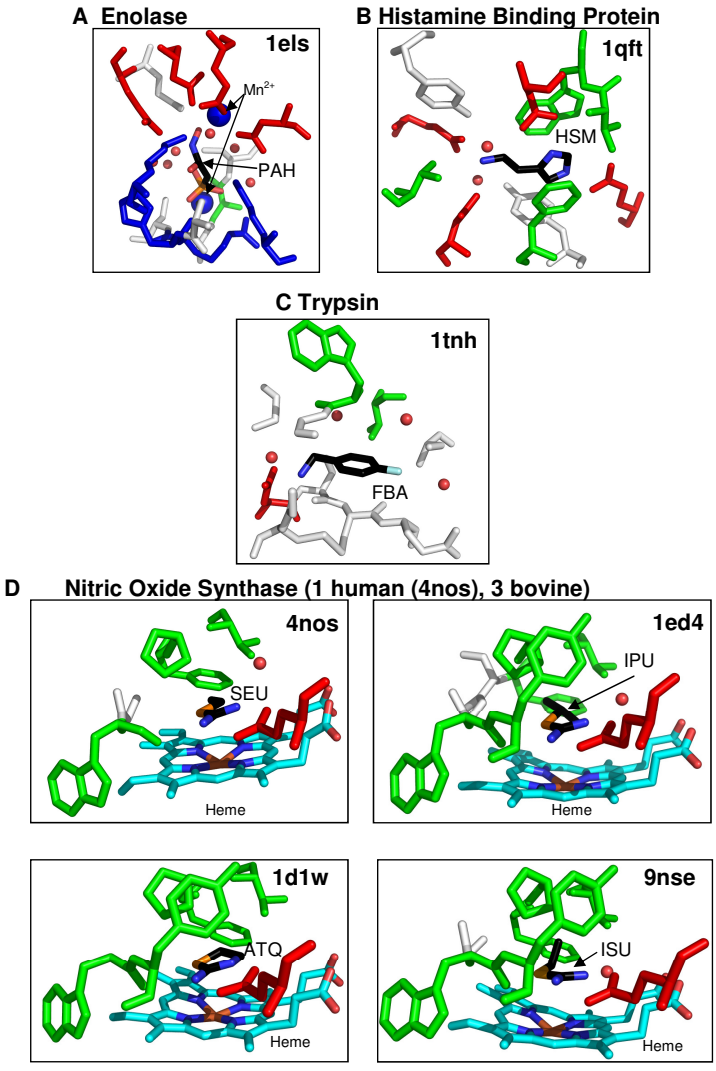
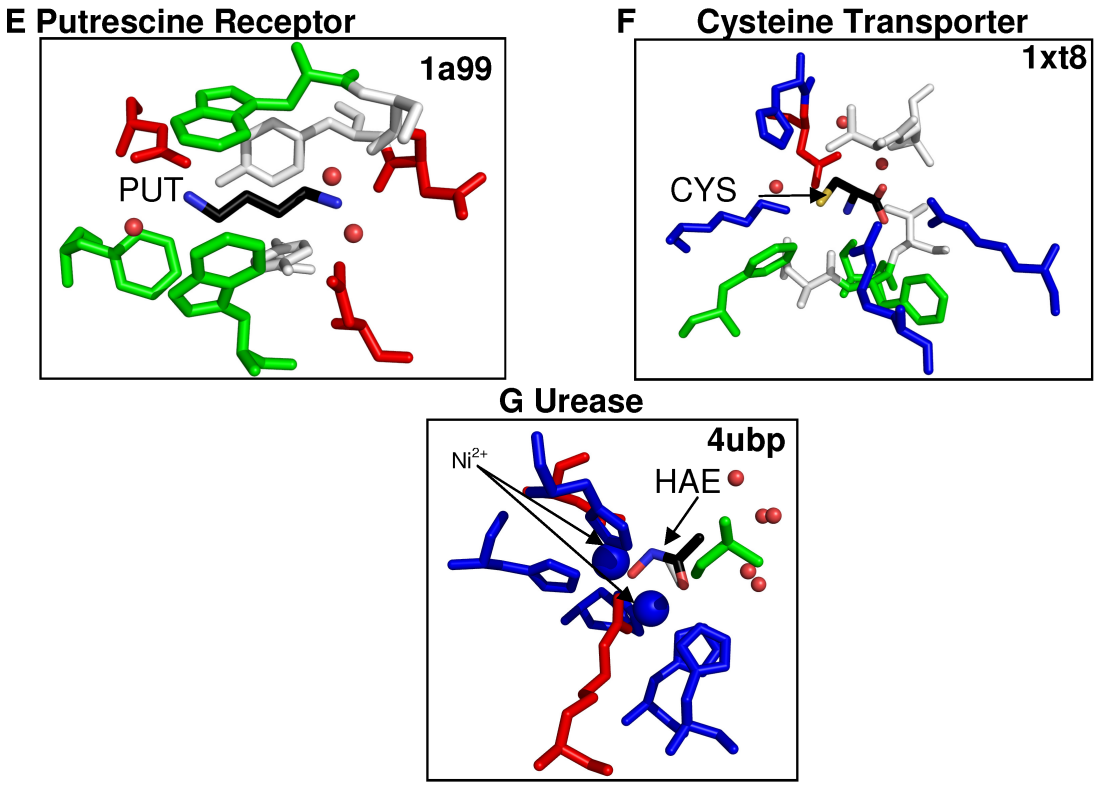


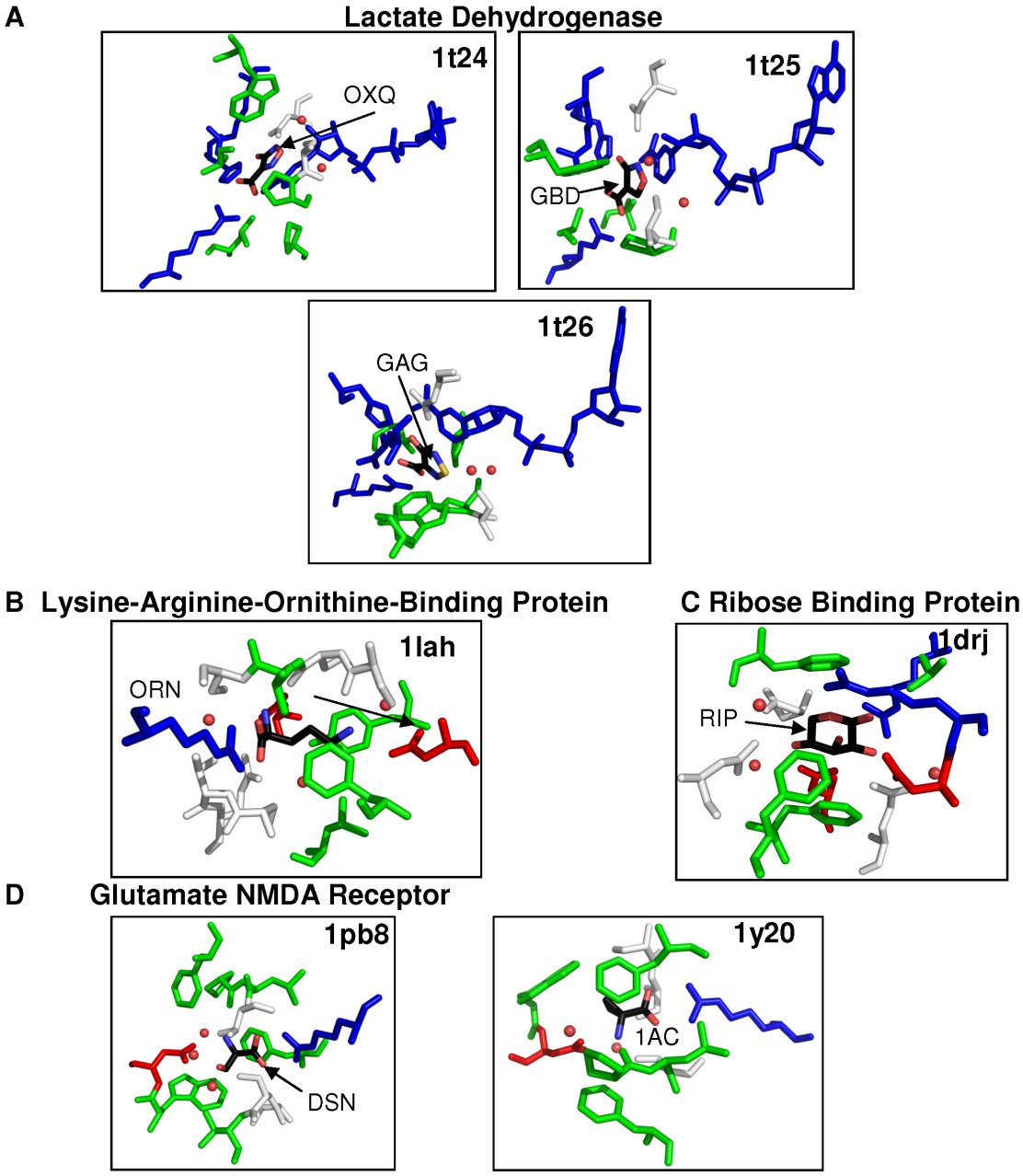
Figure 5.4: Figure 5.3 continued.



a charged ligand, see Figure 5.5. Three are lactate dehydrogenase bound to singly-charged, azol-based carboxylic acids (1t24, 1t25, and 1t26) (225). Other complexes include 1lah (an ornithine with three charge sites bound to the lysine-argninine-ornithine-binding protein) (226) and 1y20 and 1pb8 (the glutamate NMDA receptor binding a zwitterion and D-serine, respectively) (227; 228). The only structure without a charged ligand is ribose bound to D-ribose-binding protein (1drj) (229). Although the ligand is not charged, the binding site in this structure contains four charged residues (plus two asparagines and a glutamine) each making hydrogen bonds with the ribose (Figure 5.5E).

Based on the known size dependence of ligand efficiency, it is not surprising that the best ligands are small. However, the ten most efficient complexes in 5.6 are still the ten most efficient when scaled to counter small-size artifacts as suggested by Reynolds

Figure 5.5: Binding sites of highly efficient complexes (affinity per buried cavity surface area). Figures show all residues within 4 $\mathring{A}$ of the small molecule ligand. The ligand is colored by atom type. Acidic residues are colored red; basic residues are colored blue; hydrophobic residues are colored green; and hydrophilic residues are colored white as in figure 5.3. The NAD+ of lactate dehydrogenase is colored blue because the moiety against the ligand is positively charged. Water is colored red and shown in small spheres.



et al. (208). Also, it is important to note that not all small charged molecules in charged binding sites are highly efficient. If we focus on all highly charged ligands that contain 5-10 heavy atoms, we see that there are several small molecules that have more modest efficiencies (-0.40 kcal/mol-atom or poorer). Note that we are using a rather high cutoff to define less efficient binding as drugs often have efficiencies this high (a 1-nM ligand with $\geq$31 non-hydrogen atoms has $\geq$400 MW and an efficiency of -0.4 kcal/mol-atom or less). Higher efficiency cutoffs have been recommended for small molecules (207).

To determine why these ligands are less efficient, we examined all ligands with 5-10 non-hydrogen atoms and more than one charged site (63 complexes, of which 9 have efficiencies of -0.4 kcal/mol or weaker). We used two metrics to determine a complementary fit between the ligand and protein. First, the average distance between charged groups of the protein and those of the ligand was used to calculate the degree of complementarity in a way that is independent of the number of charge sites in the ligands and pockets. Second, we calculated the exposed surface area and normalized for the number of non-hydrogen atoms (ESA/size). Figure 5.6 presents the relationship of efficiency to those metrics. There is a very significant difference (two-sided Wilcoxon p-value = 0.005) in the efficiencies of complexes that are well buried (ESA/size $<$ 2 Å^2/atom) versus those that are more exposed, Figure 5.6A. The median efficiency of well-buried ligands is -0.83 kcal/mol-atom versus a median efficiency of -0.57 kcal/mol-atom for those with ESA/size $>$ 2 Å^2/atom (mean efficiencies are -0.81 versus -0.60 kcal/mol-atom, respectively). Furthermore, if those efficiencies are compared to the average distance between charged groups (Figure 5.6B), it appears that longer distances severely limit the maximum efficiency possible for the system. This is in keeping with Nissink's proposal that maximal efficiency should be proportional to contacts normalized for ligand size (206).

It should be noted that there are five systems that are not included in Figure

5.6 because they do not fit our definition of multiply charged. Though each has two titratable, all include an amine that is tightly coordinated to a metal cofactor, making it neutral. These systems have very short average contact distances, but very poor ligand efficiencies. The binding event must include a change in ionization, which is unfavorable and leads to reduced binding. To avoid confusion, these have been excluded.

For every increase of 1 Å in the average contact distance, the maximum efficiency drops by 0.7 kcal/mol-atom. Perhaps a more appropriate view is that a ligand's maximum efficiency is reduced by 0.1 kcal/mol-atom for a misfit as small as 0.14 Å in the average contacts between its charged groups and the protein's. Such significant gains/losses for such small spatial changes in the charges may explain why synthetic modifications to ligands that alter polarization and charge distribution can be so effective. The importance of charge interactions may support the ideas of optimizing charge complementarity that has been developed by Tidor and co-workers (28; 30; 230). They developed an analytical solution the Poisson-Boltzmann equation to model the electrostatics of the binding site and an analytical method of optimizing the charge profile of the ligand to match the calculated electrostatics of the binding site while also accounting for the desolvation penalty (28; 30; 230).

The importance of charge complementarity in ligand binding can be supported by other biological binding events. The ability of salt bridges to improve the stability of protein-protein interactions in protein folding or protein-protein binding may be supportive (231). Networks of salt bridges have been shown to stabilize proteins, although the majority of individual salt bridges have been shown to be destabilizing in proteins (232; 231). In a statistical study of 94 proteins from the PDB, Musafia *et al.* found that one-third of all residues participating in salt-bridges were involved in 'complex' salt bridges, which they defined as ones involving three or more amino acids (233). Olson *et al.*, were able to stabilize α-helical peptides by engineering multiple

Figure 5.6: Relationship between efficiency, exposure, and protein contacts for ligands with 5-10 atoms and more than one charge site. (A) The distribution of efficiencies is compared for systems with well buried (black) versus more exposed sites (white); a cutoff of 2 Å^2/atom is used to define the two sets. (B) Efficiencies are compared to the average contact distance between charged groups (black circles denote systems with ESA/size < 2 Å^2/atom, and white circles are ESA/size > 2 Å^2/atom). The line highlights the drop in maximal efficiency as the contacts become less favorable: roughly 0.7 kcal/mol-atom for every 1 Å increase in the average contact distance. The gray background notes systems with more modest efficiencies. The error bar indicates the standard deviation of the average of two affinity values reported in the literatures (1; 2).

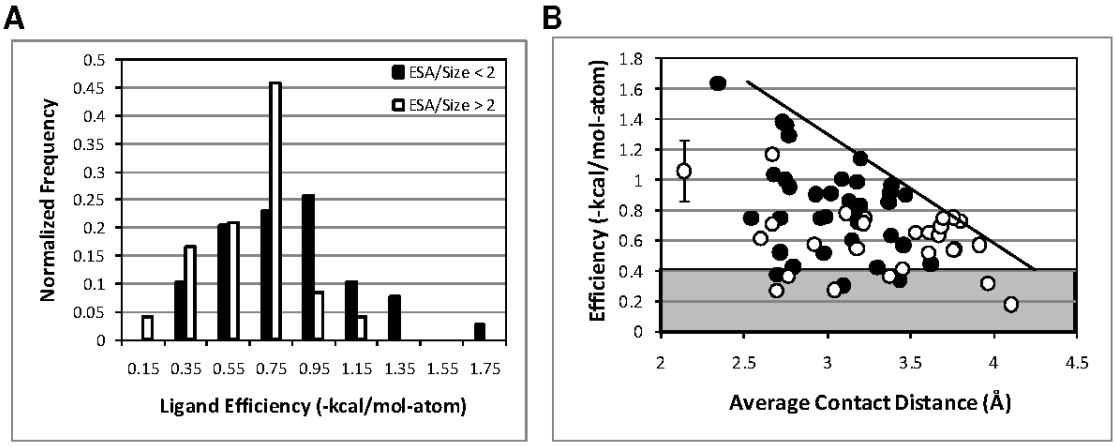


Table 5.1: Properties of small charged ligands, all ligands are between five and ten heavy atoms. Efficiency is affinity/size.

PDB Id Distance	Ligand	Affinity (-kcal/mol)	Efficiency (-kcal/mol-atom)	ESA/size (Å^2)	Avg
1a99	PUT	7.74	1.290	0.12	2.76
1ahy	MAE	4.56	0.570	0.41	3.45
1amk	PGA	5.84	0.649	3.16	3.53
1b74	DGN	1.77	0.177	8.36	4.10
1cea	ACA	6.74	0.749	10.41	3.22
1czc	GUA	3.84	0.426	0.27	2.79
1cze	SIN	3.56	0.445	0.49	3.62
1ebg	PHA	14.71	1.63	0	2.34
1egh	PGA	7.74	0.860	0.93	3.13
1el5	DMG	2.39	0.341	0.07	3.43
1ftj	GLU	8.27	0.827	0.13	3.20
1ii5	GLU	8.99	0.894	1.52	3.47
1kc7	PPR	7.50	0.750	7.46	3.75
1kv5	PGA	5.74	0.637	3.54	3.67
1lah	ORN	10.22	1.14	0.10	3.19
1m1b	SPV	6.33	0.633	0.085	3.38
1o4m	MLA	1.90	0.271	9.85	2.69
1o4n	OXD	1.90	0.317	7.93	3.96
1pb8	DSN	7.00	1.000	0	2.74
1pot	SPD	7.47	0.747	0.02	2.96
1poy	SPD	7.47	0.747	0.03	2.54
1qds	PGA	5.87	0.652	3.47	3.61
1s89	PGA	7.11	0.790	1.27	3.17
1s8a	PGA	5.43	0.600	0.93	3.15
1ssq	CYS	8.15	1.16	2.43	2.66
1tok	MAE	3.35	0.419	0.23	3.30
1txf	GLU	9.84	0.984	0.56	3.17
1usk	LEU	8.69	0.966	0.01	3.38
1wdn	GLN	9.51	0.951	0.17	2.77
1xt8	CYS	9.51	1.36	0.07	2.75
1y1m	AC5	2.47	0.274	1.62	3.04
1y1z	192	4.19	0.523	0.14	2.71
1y20	1AC	7.23	1.03	0.02	2.67
1z16	LEU	7.66	0.851	0.02	3.36
1z17	ILE	8.21	0.913	0	3.37

Table 5.2: Table 5.1 Continued

PDB Id Distance	Ligand	Affinity (-kcal/mol)	Efficiency (-kcal/mol-atom)	ESA/size (Å^2)	Avg
1z18	VAL	7.33	0.917	0.01	3.38
1zha	PEP	10.05	1.01	1.27	3.09
2aay	GPJ	3.76	0.376	0.10	2.69
2dua	OXL	3.67	0.611	5.56	2.60
2fpz	270	5.38	0.538	3.68	3.75
2gga	GPJ	3.02	0.302	0.10	3.09
2ggd	GPJ	5.16	0.516	0.08	2.98
2iqd	LPA	6.24	0.780	8.21	3.11
2o1c	PPV	6.41	0.713	8.71	3.21
2pt9	2MH	6.05	0.757	0.45	2.99
2pyy	GLU	9.10	0.910	0.22	3.02
2qrl	OGA	5.43	0.543	10.63	3.17
2rk7	OXL	5.50	0.917	6.01	2.13
2rke	SAT	5.55	0.693	2.86	3.68
2v2c	PGA	4.63	0.515	3.45	3.61
2v2h	PGA	5.12	0.569	3.12	3.91
2v7x	MET	6.48	0.720	0.002	3.18
2ypi	PGA	6.55	0.728	3.38	3.79
2ze3	AKG	4.14	0.414	3.04	3.45
2zlz	GLU	3.64	0.364	7.98	3.37
3bm5	CYS	4.02	0.574	1.88	2.92
3bra	AEF	3.67	0.367	10.80	2.76
3bu1	HSM	11.07	1.38	1.60	2.73
3bxe	13P	7.44	0.744	3.08	3.69
3epa	PUT	4.48	0.748	0.55	2.71
3epb	PUT	5.43	0.906	0.33	2.93
3jdw	ORN	4.89	0.543	1.30	3.76
3kiv	ACA	6.38	0.709	12.50	2.67

salt bridges, and found that the amount of stability obtained was cooperative (234). Networks of salt bridges were also found to be stabilizing by Kumar and Nussinov using continuum electrostatics to computationally determine the difference in energy of salt bridges compared to their hydrophobic isosteres, where the partial charges on the residue were set to zero (235). They found that the stability for most salt bridges was determined largely by the desolvation penalty; however, the networked salt bridges were an exception to this phenomenon. In all cases, the networked salt bridge was found to be stabilizing, despite a large desolvation penalty (235). These networked salt bridges are homologous to our charged ligands complemented by multiple charged residues in their binding sites.

Furthermore, Having a higher charge has been noted to be beneficial in metal ion binding to DNA/RNA. In these cases the dicationic Mg^{2+} is the preferred counter ion, compared to Na^{+}, for binding to and stabilizing the phosphate backbone of the nucleic acid (236).

Coulombic forces are the strongest non-bonded interactions that can be made, and it may not be surprising that highly efficient molecules utilize the strongest forces per atom. However, it is surprising that the free energy of binding is high since the desolvation penalty for charged molecules is insignificant (20). Since these also bind with relatively high affinity, the penalty must not be as large as previously thought. In support of the idea that the desolvation penalty is less, almost all of the structures contain water in the binding sites (Figures 5.3 and 5.5), so not all of the water are displaced.

A possible reason the desolvation penalty may be lower than initially thought is that water cannot completely solvate the charges. Many of the systems have ligands and pockets with charges that are closely spaced - too close for water to pack around each charge independently. It has been shown that the multiply charged phosphate backbone of DNA, which puts charges close together, leads to "frustrated

water" around the DNA. The restructuring of water was determined to dominate the interaction of polyols with DNA (237).

The limits of efficiency may be set by closely packed charged molecules because we are approaching covalent bonding. Zhang and Houk investigated 1017 enzymes-transition state complexes as well as 160 enzyme/inhibitor complexes. They found that transition states, which tend to have covalent or partially covalent bonds to the protein, had affinities of $K_a = 10^{16} M^{-1}$, while the inhibitors only bound with $K_a = 10^9 M^{-1}$. Additionally, they proposed that any enzyme proficiencies and affinities of greater than 10^{11} M^{-1} (˜15 kcal/mol) would exhibit covalent or partial covalent bonding (34; 213). At heavy atom distances less than 2.5 Å, low barrier hydrogen bonds exhibit at least a partial covalent nature, and provide stability of 10-20 kcal/mol (238; 239). Additionally, metals have the ability to exhibit coordinate-covalent bonding to ligands (240). In a few highly efficient complexes, we observe distances less than 2.5 Å between atoms capable of hydrogen bonding, and some cases have metals involved in coordinating the ligand. We should note that we do not believe these systems to be overly influenced by partial bonding characteristics because all are reversibly bound, many with affinities in the μM and nM range. furthermore, in the NOS system (4nos, 1ed4, 1d1w, and 9nse) where the small molecule is near a heme, the distances to the iron are greater than 4 Å. Also, investigation of the available electron densities does not indicate partial bonding between the heme and ligand.

5.3.3 Maximum affinity of ligands

What defines the maximum binding affinity of ligands? Kuntz *et al.* found that binding affinity plateaus after ˜15 atoms and little improvement is seen for larger ligands (27). No ligand has a binding affinity of -20 kcal/mol or better. In fact, it is rare to exceed -15 kcal/mol (0.1% of the complexes in Figure 5.1). Kuntz and coworkers suggested that other biological factors may be the cause of the limit; for

instance, molecules with too high of a binding affinity can exhibit clearance problems in the body (27). Nature would tend to disfavor such molecules.

Kuntz notes that affinities better than -15 kcal/mol are so tight that a ligand will most likely never dissociate before the protein is degraded. If we assume that k_{on} is the rate of diffusion of ˜$10^6 M^{-1}s^{-1}$, then an affinity of -15 kcal/mol ($K_d \approx$ 10 pM) would correspond to an average bound lifetime of ˜1 day, which is well within the lifetime of most proteins (241), but at -16 or -18 kcal/mol, the lifetimes would be approximately 6 and 187 days, respectively. However, we do not agree that clearance issues limit binding because protein binding predates complex organisms. Instead, we hypothesize that once a ligand is bound for the lifetime of a protein, there is no evolutionary pressure to coax ligands and proteins to associate more tightly.

Reynolds *et al.* have also discussed the plateau at -15 kcal/mol (208). They noted that as size increased, the maximal efficiency would decrease. They suggested the reason for the drop in efficiency was that larger ligands would need to optimize a larger number of contacts with the protein that would lead to structural compromises and thus a reduced affinity (208). We acknowledge that our data could also support this proposal because significant drops in efficiency can come from rather minor misfits in charge complementarity.

Several other factors may also contribute to the -15 kcal/mol limit to binding. First, assays which are used to determine binding constants have inherent limitations when measuring high affinity. We do not believe that this is the cause of the limit. If it was the cause, the distribution of binding affinities would drop off rapidly as one approaches the limit, which is not the case. The distributions in MOAD folow a near-normal distribution with centered at ≈9 kcal/mol. Second, our study has the limitation of examining only proteins and ligands that can be crystallized. Given that higher affinity complexes are generally hydrophobic than lower affinity complexes (210), solubility issues may limit the crystallization process. Therefore, these

structures could be under represented in our dataset. Third, most of the high affinity complexes are man-made compounds. In the drug design process, once one obtains a small molecule that binds well enough, there is no need to synthesize tighter binding small molecules. In fact ADME/Tox issues may discourage pursuing molecules in this range. Lastly, some affinities of greater than -15 kcal/mol may be incorrectly considered covalent (34; 213). We may see a limit because Binding MOAD does not contain covalently bound ligands.

5.4 Conclusions

The difficulty in determining which interactions dominate the contribution to the free energy of binding has limited the ability of researchers to predict *a priori* which small molecules will bind to a target and how tightly. Previously, it had been suggested that van der Waals and hydrophobic interactions were the driving force for small molecule binding (4; 27). Our study and other recent studies have pointed to the importance of electrostatics in driving these interactions (28; 230; 235; 30; 233; 234). We have looked at the most efficient protein-ligand complexes and have noted that in all of these complexes the small molecules have at least one charge-charge interaction, and several of them have multiple charge interactions. We highlight the importance of not only matching the shape of the binding pocket, but also complementing the charge profile of the active site.

Although desolvation of charged molecules is a barrier to binding, it appears that the small size and close proximity of charges leads to water's inability to fully solvate the ligand and its binding site. Desolvation of the charged pocket may not be as difficult to overcome, and many of the systems examined here retain some water in their sites.

Lastly, we suggest that the ≈15 kcal/mol limit of binding may be due to the fact that there is no evolutionary pressure to create tighter binding small molecules once

the bound lifetime exceeds the lifetime of the protein.

CHAPTER VI

Conclusion

We have created and utilized Binding MOAD to investigate the biophysical properties with which proteins bind to ligands. We have also committed to annual updates of the dataset to keep pace with the growth in the PDB. Binding MOAD has over thirteen thousand, hand-curated, protein-crystal structures that contain biologically relevant ligands. Binding affinity data is available for almost one-third of the entries. In the future, we wish to contain more binding-affinity data (including the addition of K_M for cofactors). Part of the value of Binding MOAD is in its careful curating and in its size and wealth of data.

Binding MOAD has plans for even greater improvement. We will add similarity-based searches for the ligands. Furthermore, we have been able to use text-mining tools to speed up our annotation process, and we are looking to make these tools, such as BUDA, which was developed in conjunction with Torrey Path, available online. This will allow users to mine text for additional types of data. Natural language processing (NLP) is proving to be a valuable tool in aiding the curation of Binding MOAD. It has significantly sped up the process of the annual updates of adding data. Such NLP-based, text-mining approaches can be readily applied to other bioinformatic projects. This technology can be used to extract a wide variety of data - not just binding information - from the huge body of literature available

today.

From chapter three we find that most ligands are well buried. This fits the common paradigm that many contacts between the ligand and the protein are a significant factor in the specificity of molecular recognition. Since most of our sites are highly buried, the majority of the cavity surface is defined only by contacts to the protein. This typically makes the portion of the surface defined by the enlarged ligand surface (ELS) only a small percentage. Figure 3.9 shows that the largest ligands tend to have more exposed surface area. These large ligands are typically peptide, nucleic acid, or sugar chains, and one would expect the patterns of binding such molecules to start to resemble the patterns of proteins binding macromolecules, such as other proteins or DNA.

Future efforts with Binding MOAD will allow us to compare broadly the binding affinity data to the patterns of molecular recognition mined from the PDB. Past studies have mined subsets of the PDB with various structural analyses of proteins and ligands (49; 154; 59; 163; 136; 157; 164; 158; 165; 159; 160; 161; 156; 142; 17; 162; 132; 166; 149; 167), but now, we will be able to add another layer of depth to such studies. There is more to binding affinity than just burying a ligand inside a protein, and all of the complex issues that go into creating an effective scoring function (168) will need to be considered in future analyses. Both shape and chemical complementarity are thought to be the basis of molecular recognition. Our future analyses will have to consider the chemical complementarity or what "types" of surfaces are solvent-exposed or interact with the protein, in order to understand how to improve the enthalpy of binding. We will also need to address the very complex issue of entropic changes upon binding.

In chapter four of the thesis, physical characteristics of bound ligands in Binding MOAD were compared between enzymes and non-enzymes as well as high-affinity and low-affinity complexes. The comparison between ligand sizes for low-affinity versus

high-affinity binding shows that divergent approaches are likely needed to improve the affinity of enzyme inhibitors versus those for non-enzymes. The traditional approach of adding functional groups to fill more of the pocket may work for enzymes, but it may not be as appropriate for non-enzyme systems. However, making ligands more hydrophobic appears to aid binding in both enzymes and non-enzymes.

Non-enzymes have higher ligand efficiencies than enzymes, which may be a reflection of their biological roles. This is also encouraging when considering the druggability of non-enzymes. The differences in efficiencies between enzymes and non-enzymes could not be attributed to the small molecules or the protein alone. Therefore, future investigations would require one to look at the specific contacts between the protein and ligand or entropic considerations.

In the pharmaceutical industry, ligand efficiencies have become a metric for evaluating hits from screening campaigns and even candidate compounds.(25) Our results would caution against applying a rigid standard across all protein targets, since each individual protein family showed different ligand efficiencies. At the very least, a cutoff based on ligand efficiency should differ between enzymes and non-enzymes. Ideally, cutoffs would differ between protein families and only be considered as one of several guidelines in a selection process.

We have also noted that Binding MOAD provides strong support of several mathematical models cited above,(199; 185; 176) particularly those of Hajduk and coworkers. Our results have implications for the development of scoring functions for docking and predicting druggability of a binding site.(200; 201; 202; 203) The differences between non-enzymes and enzymes, as well as the differences across enzymatic systems, underscore the challenges of developing universal functions that perform well across all systems. Modest improvement might be achieved by developing separate functions for enzymes and non-enzymes, with even greater improvement expected for functions trained on specific protein families.

In chapter five, we have looked at the most efficient protein-ligand complexes in Binding MOAD and have noted that in all of these complexes the small molecules have at least one charge-charge interaction, and several of them have multiple charge interactions. We highlight the importance of not only matching the shape of the binding pocket, but also complementing the charge profile of the active site.

Although desolvation of charged molecules is a barrier to binding, it appears that the small size and close proximity of charges leads to water's inability to fully solvate the ligand and its binding site. Desolvation of the charged pocket may not be as difficult to overcome, and many of the systems examined here retain some water in their sites. Future work to help understand water's role may be to use isothermal calorimetry to provide entropic and enthalpic contributions to the free energy of binding. Additionally, we may be able to use quantum mechanics and molecular mechanics to investigate the energy of solvation for these highly electrostatic ligands and binding sites and how it compares to the free energy of solvation of other highly electrostatic ligands and binding sites which do not bind with high efficiency.

It is also important to note that all of our results are developed from structures that can be crystallized. Since our dataset is significantly large, we believe our results to be applicable. However, it will be important in the future to compare the types of small molecules and proteins in Binding MOAD to the entire space of drug-like small molecules and protein binding sites, to see if this is true representation of all complexes. Although no crystal structures are available for all protein-ligand complexes, it would still be possible to compare the chemical properties of the small molecule and the amino acid compositions of the proteins. This will be a daunting task as it will require a dataset much larger than that presented in Binding MOAD.

APPENDICES

APPENDIX A

Supplemental Information for Chapter 4

A.1 Distributions, box plots, and distribution analysis

see Figures A.1-A.7

Figure A.1: This figure shows the relevant statistical figures regarding the distribution of size (a_heavy) in heavy atoms for the four classifications.

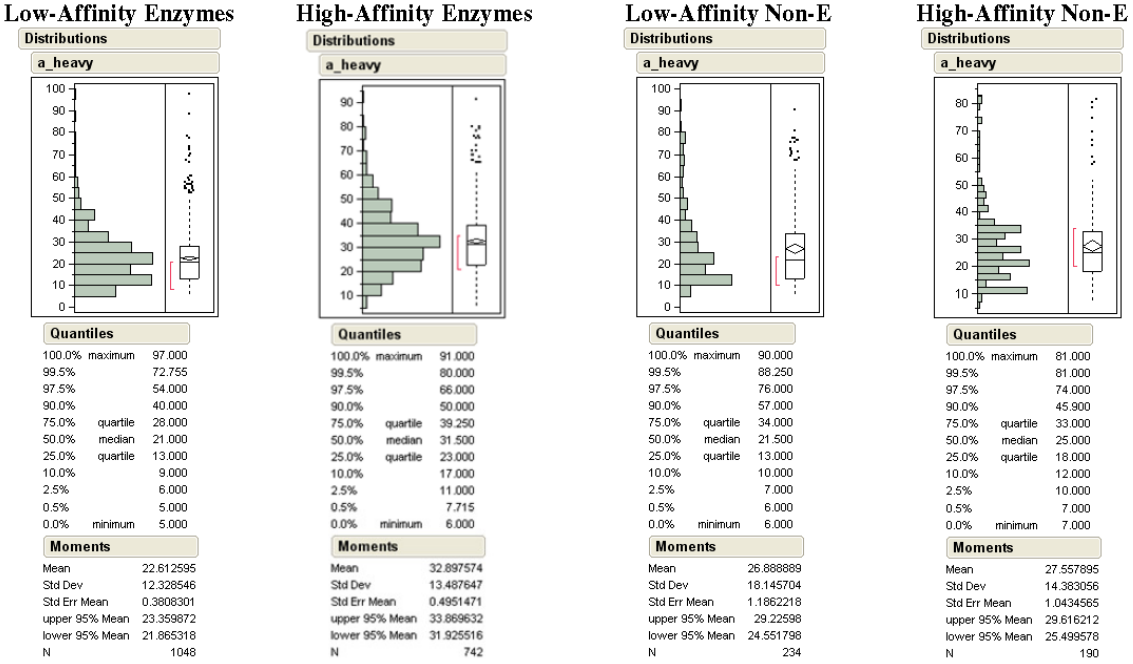


Figure A.2: This figure shows the relevant statistical figures regarding the distribution of BSA ($\AA^2$) for the four classifications.

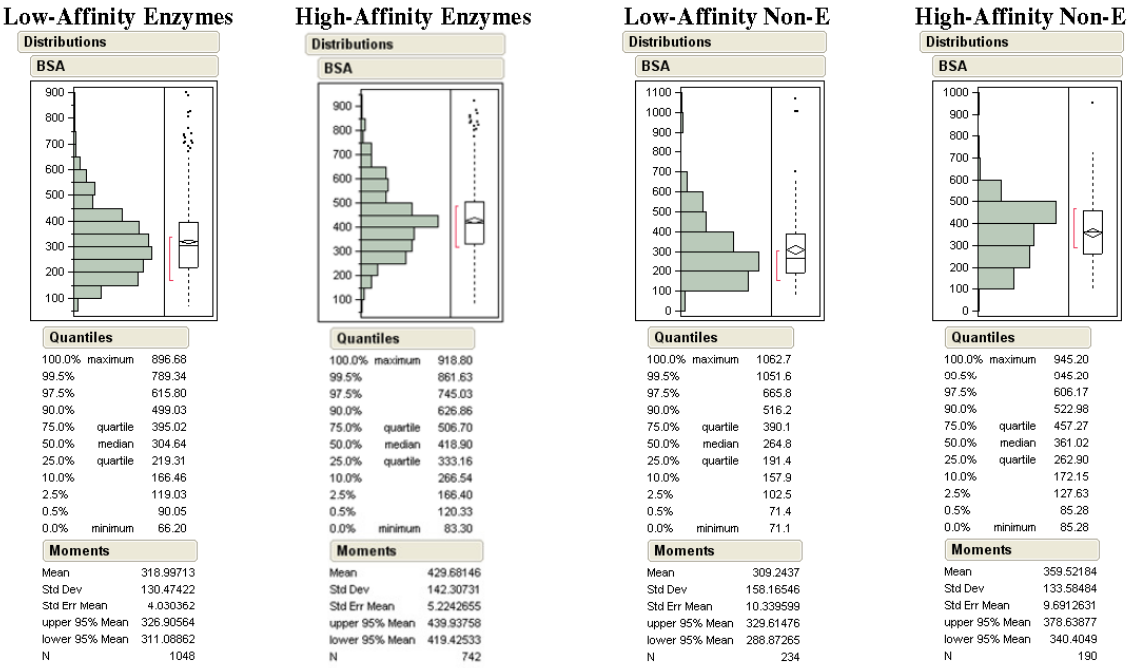


Figure A.3: This figure shows the relevant statistical figures regarding the distribution of ESA ($\AA^2$) for the four classifications.

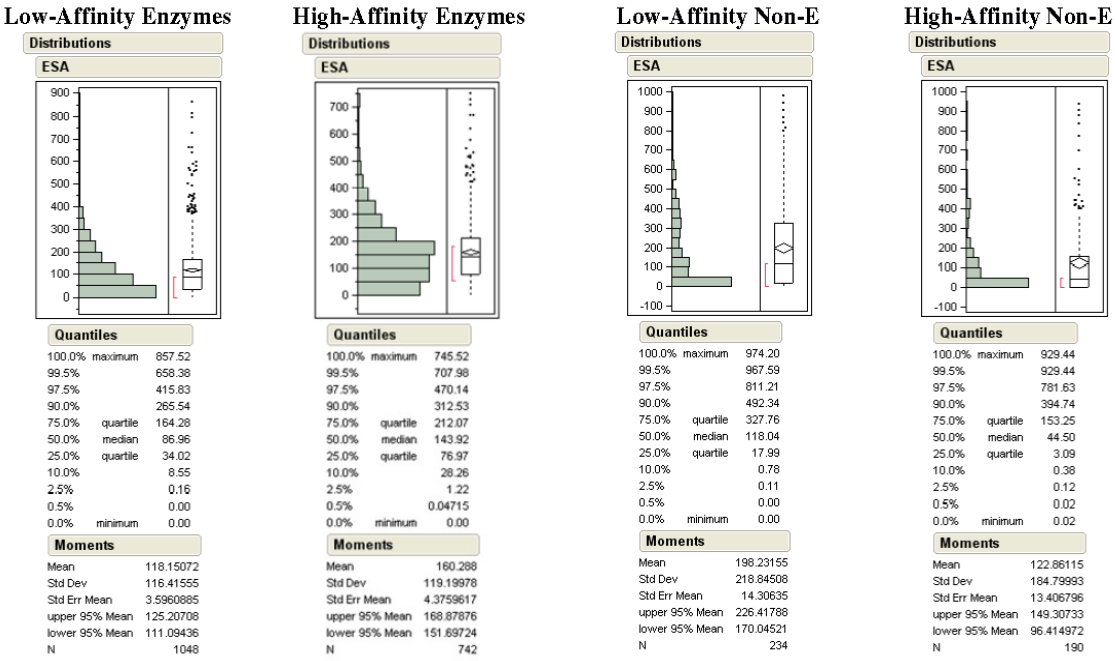


Figure A.4: This figure shows the relevant statistical figures regarding the distribution of sqrt(ESA)(Å) for the four classifications.

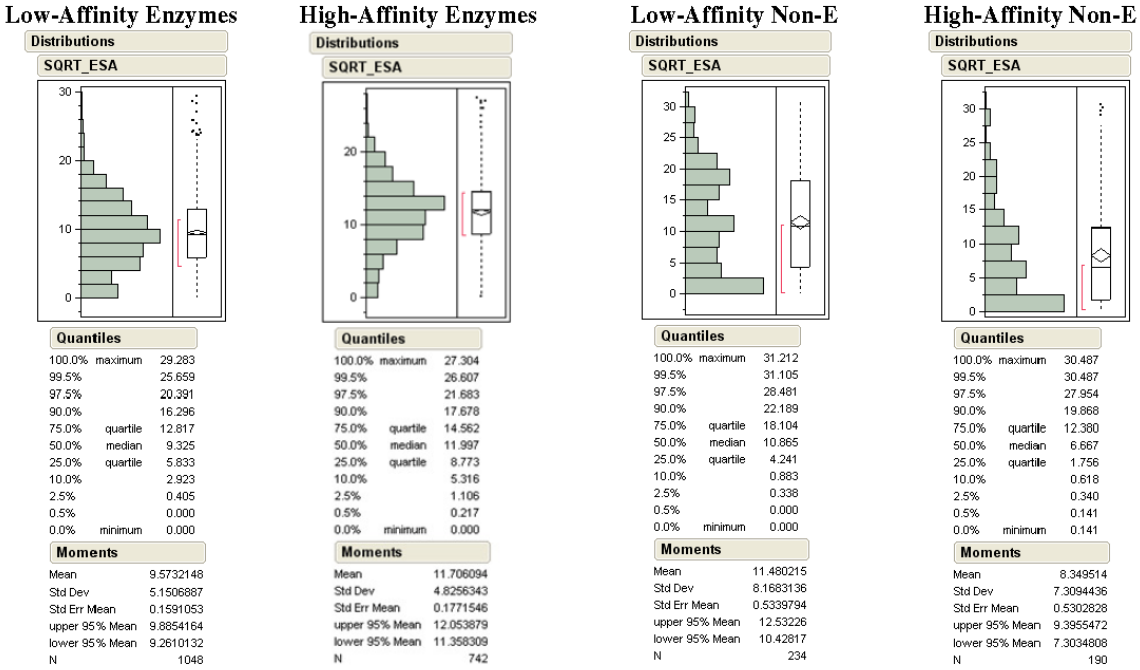


Figure A.5: This figure shows the relevant statistical figures regarding the distribution of size ligand efficiency (kcal/mol-atom) for the four classifications.

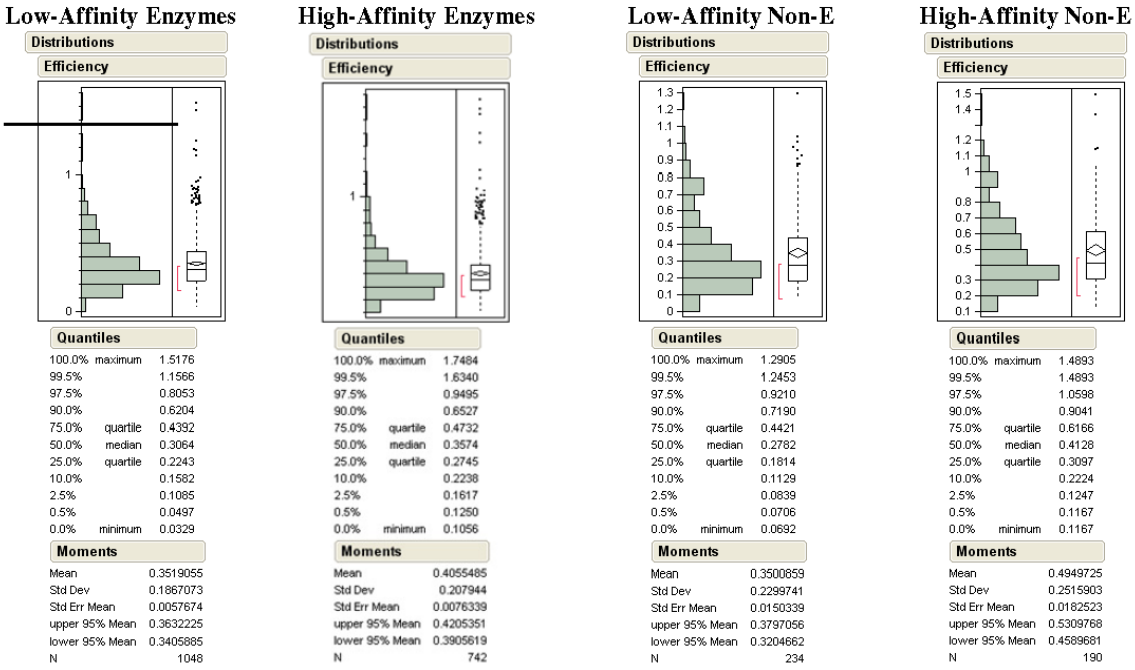


Figure A.6: This figure shows the relevant statistical figures regarding the distribution of BSA ligand efficiency (cal/mol-$\mathring{A}^2$) for the four classifications.

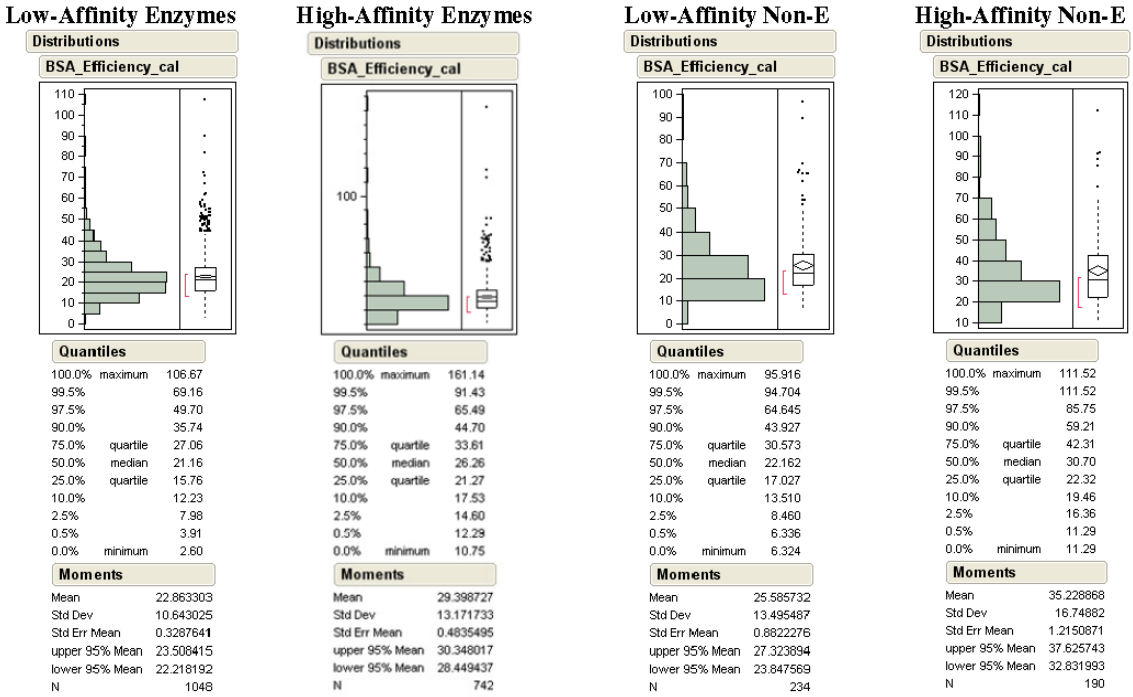
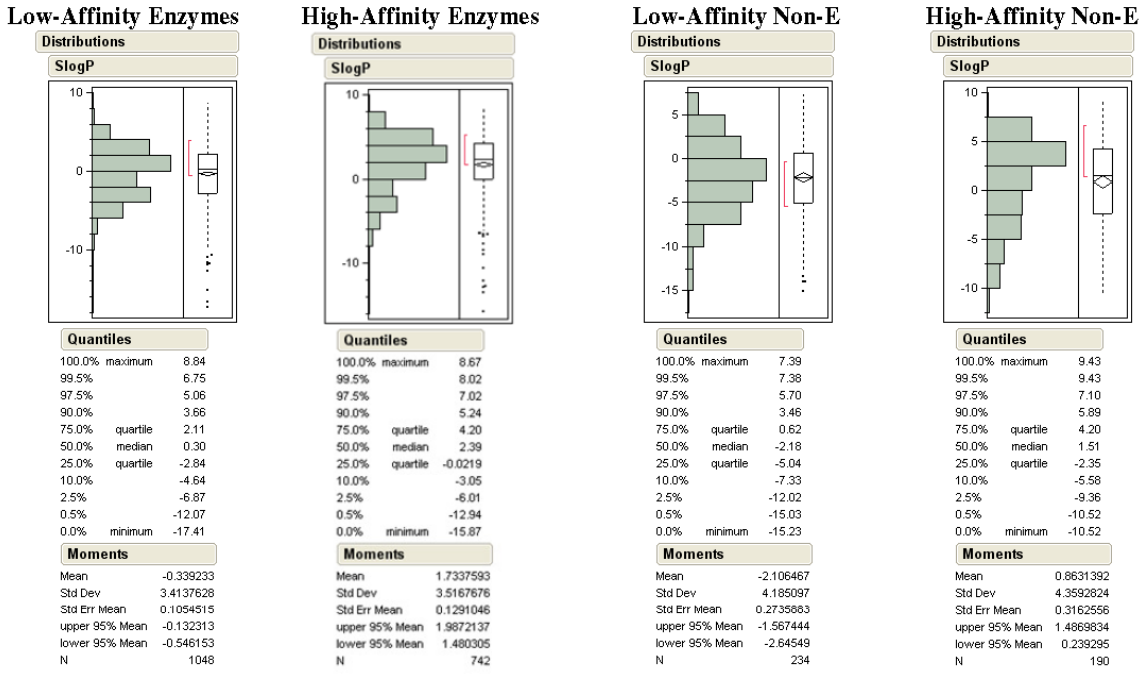


Figure A.7: This figure shows the relevant statistical figures regarding the distribution of SlogP for the four classifications.



A.2 Tukey-Kramer HSD analysis

The datasets do not have normal, Gaussian distributions, so we used square-root transformations to reduce the skew of the data. With the modification, the Tukey-Kramer HSD test could be used to examine the means (medians were compared in the p-values). The large population of the dataset also increased the significance of this statistical test. The Tukey-Kramer HSD test was performed in SAS with an exceptionally tight confidence value of 0.0001.

First, high-affinity enzymes are significantly larger than high-affinity non-enzymes and low-affinity enzymes; however, the non-enzymes are not significantly different (Table A.1). Since the size distributions are not normal we also analyzed the square-root transform of the size to shift the right-skewed distribution to a more normal distribution (Figure A.1). The same trend held when the Tukey-Kramer HSD test for this variable is performed (Table A.1).

Secondly, in the paper we noted that the high-affinity non-enzymes were significantly less exposed than the high-affinity enzymes according to the Wilcoxon test. This is confirmed with the Tukey test. The ESA distribution was not normal, and a square root transform was used to transform the ESA to a more normal distribution as well to perform the test (Figures A.2 and A.3). The Tukey grouping confirms the significance of the difference in the degree of exposure between high-affinity non-enzymes and high-affinity non-enzymes.

Third, high-affinity non-enzymes are more efficient than high-affinity non-enzymes. This can also be seen in the Tukey grouping for the two efficiency variables (Tables A.3 and A.4).

Lastly, we noted that the high-affinity complexes were more hydrophobic, according to an increase in SlogP. The high-affinity complexes group in the same group, while the low-affinity complexes group in separate groups, with the exception of the high-affinity non-enzymes and low-affinity non-enzymes which appear in the same

group.

see Tables A.1-A.5

Table A.1: This table shows the Tukey-Kramer HSD for size in number of heavy atoms (sqrt(size)) over the four classifications. Classes indicated with the same letter are not significantly different at the 99.99% confidence level.

Classification	Tukey Grouping	Means
High-Affinity Enzymes	A	32.898 (5.61754)
High-Affinity Non-Enzymes	B	27.558 (5.09263)
Low-Affinity Non-Enzymes	B C	26.889 (4.93237)
Low-Affinity Enzymes	C	22.613 (4.59402)

Table A.2: This table shows the Tukey-Kramer HSD for sqrt(ESA) over the four classifications. Classes indicated with the same letter are not significantly different at the 99.99% confidence level.

Classification	Tukey Grouping	Means
High-Affinity Enzymes	A	11.7061 Å
High-Affinity Non-Enzymes	B	8.3495 Å
Low-Affinity Non-Enzymes	A	11.4802 Å
Low-Affinity Enzymes	B	9.5732 Å

Table A.3: This table shows the Tukey-Kramer HSD for size ligand efficiency over the four classifications. Classes indicated with the same letter are not significantly different at the 99.99% confidence level.

Classification	Tukey Grouping	Means
High-Affinity Enzymes	B	0.40555 kcal/mol-atom
High-Affinity Non-Enzymes	A	0.49497 kcal/mol-atom
Low-Affinity Non-Enzymes	B	0.35009 kcal/mol-atom
Low-Affinity Enzymes	B	0.35191 kcal/mol-atom

A.3 Properties after removal of Cofactors

see Table A.6

A.4 Patterns obtained from the non-redundant dataset

See Table A.7 and Figures A.8 and A.9

Figure A.8: For the non-redundant complexes: distribution of ligand sizes (number of non-hydrogen atoms) and buried surface area of the pocket (BSA in $\mathring{A}^2$) are given in normalized percent frequencies. (a) Comparisons of high-affinity complexes, (b) low-affinity complexes, (c) enzymes, and (d) non-enzymes are presented. High-affinity enzymes are shown in dark blue lines, and low-affinity enzymes are in green lines. High-affinity non-enzymes are in red, and low-affinity non-enzymes are in gold.

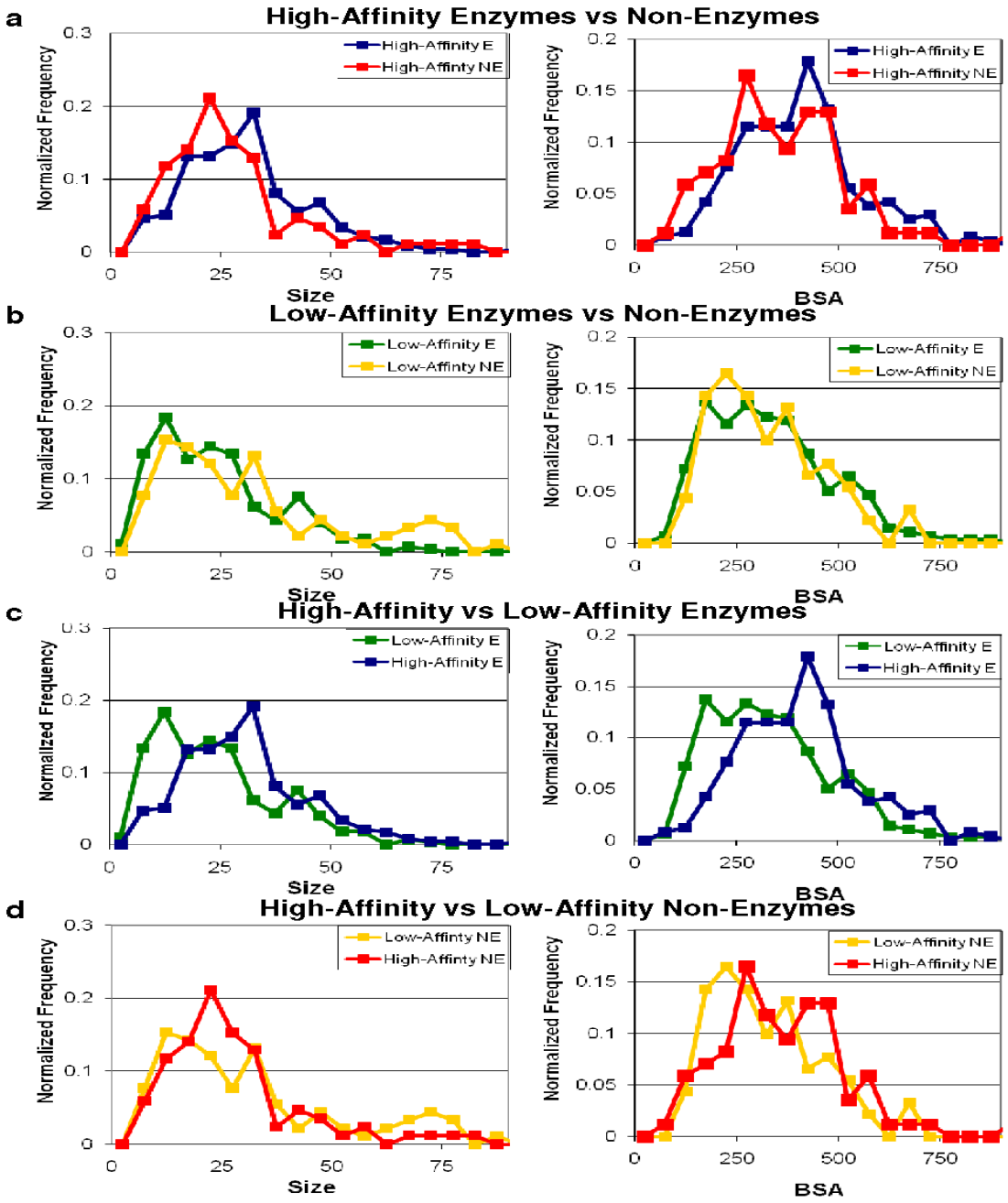


Figure A.9: For the non-redundant complexes: distribution of ligand efficiencies per size (-kcal/mol-atom) and per contact (-kcal/mol-$\AA^2$) are given in normalized percent frequencies. (a) Comparisons of high-affinity complexes and (b) low-affinity complexes are presented. High-affinity enzymes are shown in dark blue lines, and high-affinity non-enzymes are in red lines. Low-affinity enzymes are in green lines, and low-affinity non-enzymes are in gold lines.

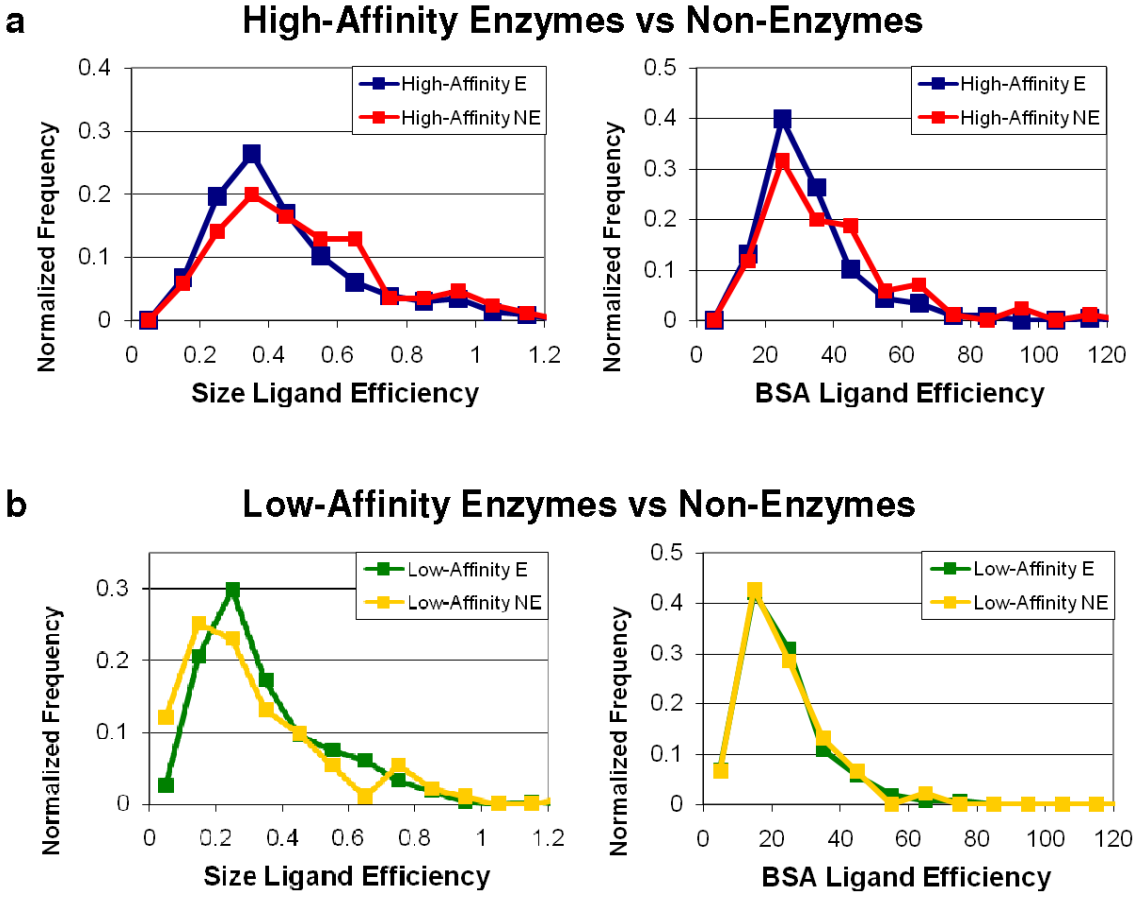


Table A.4: This table shows the Tukey-Kramer HSD for BSA ligand efficiency over the four classifications. Classes indicated with the same letter are not significantly different at the 99.99% confidence level.

Classification	Tukey Grouping	Means
High-Affinity Enzymes	B	29 cal/mol-$\mathring{A}^2$
High-Affinity Non-Enzymes	A	35 cal/mol-$\mathring{A}^2$
Low-Affinity Non- Enzymes	B C	26 cal/mol-$\mathring{A}^2$
Low-Affinity Enzymes	C	23 cal/mol-$\mathring{A}^2$

Table A.5: This table shows the Tukey-Kramer HSD for SlogP over the four classifications. Classes indicated with the same letter are not significantly different at the 99.99% confidence level.

Classification	Tukey Grouping	Means
High-Affinity Enzymes	A	1.7338
High-Affinity Non-Enzymes	A B	0.8631
Low-Affinity Non-Enzymes	C	-2.1065
Low-Affinity Enzymes	B	-0.3392

A.5 Patterns obtained from complexes with Kd data

See Table A.8

A.6 Three enzymes with a large range in affinities for a small range of ligand sizes

In three cases – neuramidase, MTA/SAH nucleosidase, and protocatechuate 3,4-dioxygenase – we found enzymes that had a very small range of ligand sizes and a large range in binding affinity, Figure A.10. It is unclear whether the strong overall trends of 0.62 kcal/mol-atom for neuramidase and 0.71 kcal/mol-atom for MTA/SAH nucleosidase are exceptional examples of the correlations expected for enzymes or whether they indicate that only conservative changes in sizes are allowed for these systems, as we suggested for non-enzymes. We have presented protocatechuate 3,4-dioxygenase in Figure S10c to show the only example obtained for an enzyme with overwhelmingly strong influence from small changes, much like arabinose-binding protein in the

Table A.6: Median Characteristics of Enzyme and Non-Enzyme Complexes in the Redundant Set with All Cofactors Removed from Consideration (includes K_d, K_i, and IC_{50} values for affinity). The values are nearly unchanged from Table 4.1, underscoring the robust nature of the data when 109 complexes (˜5%) are removed.

Median Physical Properties	**Low Affinity** >250 nM ΔG_{bind} > -9 kcal/mol	**High Affinity** <=250 nM ΔG_{bind} <= -9 kcal/mol
Enzymes	**972 complexes**	**715 complexes**
ΔG_{bind}	-6.6 kcal/mol	-10.9 kcal/mol
Size[a]	20 atoms	32 atoms
BSA	297 $Å^2$	420 $Å^2$
ESA (%ESA)[b]	84 $Å^2$ (21%)	142 $Å^2$ (24%)
SlogP	0.5	2.6
-ΔG_{bind}/atom	0.32 kcal/mol-atom	0.36 kcal/mol-atom
-ΔG_{bind}/BSA	21 cal/mol-$Å^2$	26 cal/mol-$Å^2$
Non-Enzymes	**231 complexes**	**187 complexes**
ΔG_{bind}	-7.2 kcal/mol	-10.5 kcal/mol
Size[a]	21 atoms	25 atoms
BSA	261 $Å^2$	358 $Å^2$
ESA (%ESA)[b]	116 $Å^2$ (32%)	40 $Å^2$ (10%)
SlogP	-2.2	1.5
-ΔG_{bind}/atom	0.28 kcal/mol-atom	0.42 kcal/mol-atom
-ΔG_{bind}/BSA	22 cal/mol-$Å^2$	31 cal/mol-$Å^2$

a. Ligand size is given in the number of non-hydrogen atoms.
b. Percent exposure is ESA/(ESA+BSA) for each individual ligand.

Table A.7: Median Characteristics of Protein-Ligand Binding in Enzymes and Non-Enzymes from the Non-Redundant Dataset[a].

	Low Affinity > 250 nM ΔG_{bind} > -9 kcal/mol	**High Affinity** <= 250 nM ΔG_{bind} <= -9 kcal/mol	**Comparison**[b]
Enzymes	**277 complexes**	**235 complexes**	High-affinity ligands are 30% larger
ΔG_{bind}	-6.8 kcal/mol	-11.1 kcal/mol	
Size[c]	23 atoms	30 atoms	
BSA	313 $Å^2$	413 $Å^2$	
ESA	93 $Å^2$ (22%)	122 $Å^2$ (22%)	
(%ESA)[d]	-1.3	1.3	
SlogP	0.29 kcal/mol-atom	0.39 kcal/mol-atom	
ΔG_{bind}/atom	20 cal/mol-$Å^2$	28 cal/mol-$Å^2$	
ΔG_{bind}/BSA			
Non-Enzymes	**91 complexes**	**85 complexes**	Low-affinity ligands have more than three times the esposure
ΔG_{bind}	-6.9 kcal/mol	-10.9 kcal/mol	
Size[c]	26 atoms	24 atoms	
BSA	302 $Å^2$	343 $Å^2$	
ESA	201 $Å^2$ (43%)	64 $Å^2$ (15%)	
(%ESA)[d]	-3.2	1.4	
SlogP	0.24 kcal/mol-atom	0.44 kcal/mol-atom	
ΔG_{bind}/atom	21 cal/mol-$Å^2$	34 cal/mol-$Å^2$	
ΔG_{bind}/BSA			
Comparison[b]	Non-enzymes have similar ligand efficiencies	Non-enzymes have greater ligand efficiency	

a. Proteins are grouped by 90% sequence identity and represented by the complex with the highest affinity ligand

b. All differences noted in the comparisons sections have a statistical significance of >96% ($p<0.04$).

c. Ligand size is given in the number of non-hydrogen atoms.

d. Percent exposure is ESA/(ESA+BSA).

Table A.8: Median Characteristics of Enzyme and Non-Enzyme Complexes in the Redundant Set with Kd Values.

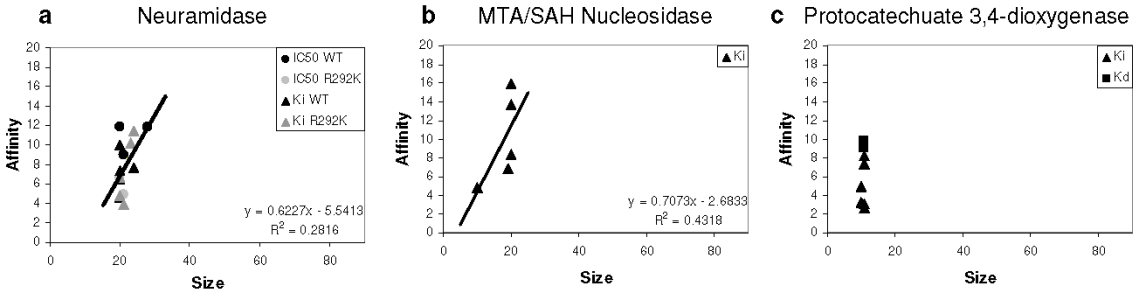
	Low Affinity > 250 nM ΔG_{bind} > -9 kcal/mol	**High Affinity** <= 250 nM ΔG_{bind} <= -9 kcal/mol	**Comparison**[a]
Enzymes	**291 complexes**	**138 complexes**	High-affinity ligands are 55% larger
ΔG_{bind}	-6.3 kcal/mol	-11.2 kcal/mol	
Size[b]	20 atoms	31 atoms	
BSA	290 $\AA^2$	345 $\AA^2$	
ESA (%ESA)[c]	90 $\AA^2$ (22%)	170 $\AA^2$ (30%)	
SlogP	-1.6	-0.21	
$-\Delta G_{bind}$/atom	0.31 kcal/mol-atom	0.39 kcal/mol-atom	
$-\Delta G_{bind}$/BSA	22 cal/mol-$\AA^2$	33 cal/mol-$\AA^2$	
Non-Enzymes	**158 complexes**	**113 complexes**	Low-affinity ligands are 2.5-3 times more exposed
ΔG_{bind}	-7.3 kcal/mol	-10.2 kcal/mol	
Size[b]	24 atoms	23 atoms	
BSA	292 $\AA^2$	342 $\AA^2$	
ESA (%ESA)[c]	131 $\AA^2$ (35%)	54 $\AA^2$ (13%)	
SlogP	-3.1	-0.09	
$-\Delta G_{bind}$/atom	0.27 kcal/mol-atom	0.44 kcal/mol-atom	
$-\Delta G_{bind}$/BSA	22 cal/mol-$\AA^2$	33 cal/mol-$\AA^2$	
Comparison[a]	Non-enzymes have similar ligand efficiencies	Non-enzymes have greater ligand efficiency	

a. All points comparing the averages have a statistical significance of 99.1% or better.
b. Ligand size is given as the number of non-hydrogen atoms.
c. Percent exposure is ESA/(ESA+BSA).

non-enzymes Figure A.10b. All the ligands for protocatechuate 3,4-dioxygenase have 10 or 11 non-hydrogen atoms, and it appears to be an example of a small, restricted binding site.

see Figure A.10

Figure A.10: Examples of enzyme families that show exceptionally strong response and limited size ranges for ligands. (a) Wild-type (K_i as black triangles, IC_{50} as black circles) and the R292K-mutant (Ki as gray triangles, IC_{50} as gray circles) of neuraminidase show the same strong response to conservative changes to the ligands. (b) Sizes and K_i (black triangles) for ligands bound to MTA/SAH Nucleosidase. (c) The data points for the ligands bound to protocatechuate 3,4-dioxygenase cannot be fit to a line because of the near vertical arrangement.



A.7 Total amino acid content in enzymes and non-enzymes

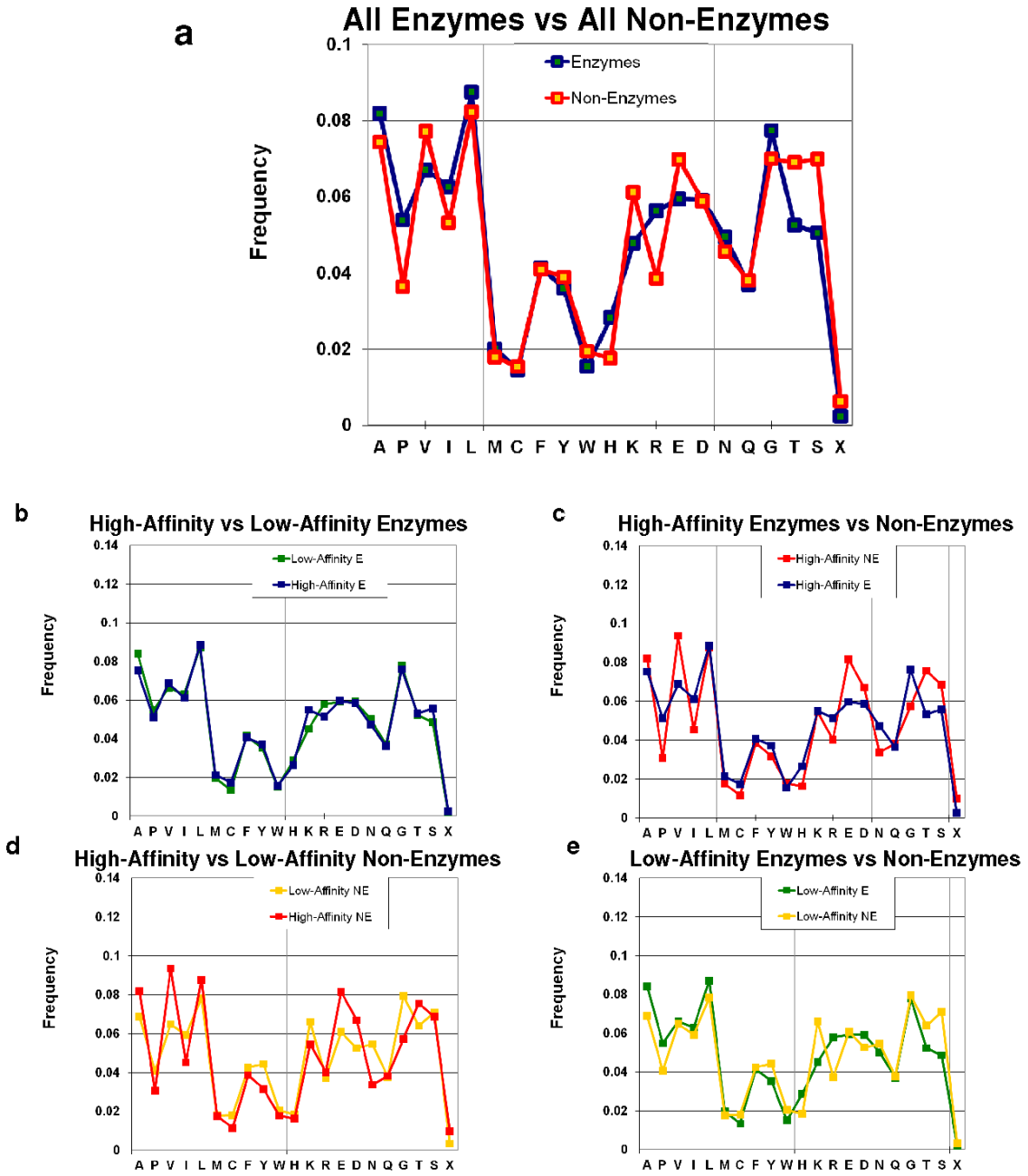
The differences in binding sites of enzymes and non-enzymes (Figure 4.6) are not due to inherent differences in the amino acid composition of the proteins.

see Figuer A.11

A.8 Classes of proteins that make up the high-affinity complexes

It should be noted that the number of enzymes and non-enzymes in the non-redundant dataset are slightly larger than the lists below. This is because a protein may be represented more than once if it comes from different species that have less

Figure A.11: Amino acid content in enzymes and non-enzymes, given in normalized percent frequencies. Amino acids are listed by hydrophobic, aromatic, cationic, anionic, and hydrophilic. "X" denotes cofactors, unnatural amino acids, and covalent modifications on the protein (does **not** include crystallographic additives in the crystal structure).



than 90% sequence identity. Any enzyme listed in the non-enzyme list is an example of allosteric binding. Though we made every effort to locate these examples, it is possible that some are contained in the enzyme set, rather than the non-enzyme set, because of the EC notation inherent to the Binding MOAD dataset.

HIGH-AFFINITY COMPLEXES OF ENZYMES ARE COMPRISED OF:

1. ALDOSE REDUCTASE
2. 1-DEOXY-D-XYLULOSE 5-PHOSPHATE REDUCTOISOMERASE
3. 2-DEHYDRO-3-DEOXYPHOSPHOOCTONATE ALDOLASE
4. 3-HYDROXYACYL-COA DEHYDROGENASE TYPE II
5. 4-HYDROXYPHENYLPYRUVATE DIOXYGENASE
6. 5'-DEOXY-5'-METHYLTHIOADENOSINE PHOSPHORYLASE
7. A/G-SPECIFIC ADENINE GLYCOSYLASE
8. ACETOLACTATE SYNTHASE, MITOCHONDRIAL
9. ACETYLCHOLINESTERASE
10. ACETYL-COENZYME A CARBOXYLASE
11. ADAM 17
12. ADENOSINE DEAMINASE
13. ADENYLOSUCCINATE SYNTHETASE
14. ADP-RIBOSYLATION FACTOR-LIKE PROTEIN 3
15. ALCOHOL DEHYDROGENASE
16. ALPHA1,2-MANNOSIDASE

17. ALPHA-AMYLASE
18. ALPHA-MANNOSIDASE II
19. ALPHA-THROMBIN
20. ANGIOTENSIN CONVERTING ENZYME
21. ARGINASE
22. ASPARTATE CARBAMOYLTRANSFERASE CATALYTIC CHAIN
23. B. ANTHRAX LETHAL FACTOR
24. BACTERIAL LEUCYL AMINOPEPTIDASE
25. BETA-1,4-XYLANASE
26. BETA-D-GLUCAN EXOHYDROLASE
27. BETA-GALACTOSIDASE
28. BETA-GLUCOSIDASE
29. BETA-SECRETASE
30. BIFUNCTIONAL DIHYDROFOLATE REDUCTASE-THYMIDYLATE SYNTHASE
31. BIFUNCTIONAL PURINE BIOSYNTHESIS PROTEIN PURH
32. CAMP-DEPENDENT PROTEIN KINASE
33. CAMP-SPECIFIC 3',5'-CYCLIC PHOSPHODIESTERASE 4B
34. CAMP-SPECIFIC 3',5'-CYCLIC PHOSPHODIESTERASE 4D
35. CARBONIC ANHYDRASE I

36. CARBONIC ANHYDRASE II

37. CARBOXYPEPTIDASE A

38. CATALASE

39. CATECHOL-O-METHYLTRANSFERASE

40. CATHEPSIN D

41. CATHEPSIN G

42. CATHEPSIN K

43. CATHEPSIN S

44. CELL DIVISION PROTEIN KINASE 2

45. CGMP-INHIBITED 3',5'-CYCLIC PHOSPHODIESTERASE 3B

46. CGMP-SPECIFIC 3',5'-CYCLIC PHOSPHODIESTERASE 5A

47. CHITOTRIOSIDASE

48. CHORISMATE SYNTHASE

49. CHYMASE

50. CITRATE SYNTHASE

51. COAGULATION FACTOR VII

52. COAGULATION FACTOR X

53. COAGULATION FACTOR XA

54. COLLAGENASE 3

55. CYTIDINE DEAMINASE

56. CYTOCHROME P450 2B4

57. CYTOCHROME P450-CAM

58. DENOSINE DEAMINASE

59. DEOXYRIBONUCLEOSIDE KINASE

60. DIHYDROFOLATE REDUCTASE

61. DIHYDROOROTATE DEHYDROGENASE

62. DIHYDROPTERIDINE REDUCTASE

63. DIPEPTIDYL PEPTIDASE IV

64. DNA GYRASE SUBUNIT B

65. ELASTASE

66. ENDOPLASMIN

67. ENDOTHIAPEPSIN

68. ENOYL-[ACYL-CARRIER PROTEIN] REDUCTASE

69. EPIDERMAL GROWTH FACTOR RECEPTOR KINASE

70. EPOXIDE HYDROLASE 2, CYTOPLASMIC

71. EPSILON-THROMBIN

72. ERK2

73. ESTROGEN SULFOTRANSFERASE

74. EXOTOXIN A

75. FERREDOXIN: NADP+ REDUCTASE

76. FIBROBLAST COLLAGENASE

77. FIBROBLAST GROWTH FACTOR (FGF) RECEPTOR 1

78. FKBP25

79. FLAVODOXIN

80. FORMYLMETHIONINE DEFORMYLASE

81. FR-1 PROTEIN

82. FRUCTOSE 1,6-BISPHOSPHATASE

83. G25K GTP-BINDING PROTEIN

84. GLUCOAMYLASE

85. GLUCOSE-6-PHOSPHATE ISOMERASE

86. GLUTATHIONE S-TRANSFERASE (ISOENZYME 3-3)

87. GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE

88. GLYCOGEN SYNTHASE KINASE-3 BETA

89. GUANINE PHOSPHORIBOSYLTRANSFERASE

90. H-2 CLASS I HISTOCOMPATIBILITY ANTIGEN, K-B ALPHA CHAIN

91. HEAT SHOCK PROTEIN 90

92. HEPATITIS C VIRUS NS5B RNA-DEPENDENT RNA POLYMERASE

93. HEPATOCYTE GROWTH FACTOR RECEPTOR

94. HISTONE DEACETYLASE 8

95. HIV-1 PROTEASE

96. HIV-2 PROTEASE
97. HMG-COA REDUCTASE
98. HUMAN BETA2 TRYPTASE
99. HYPOTHETICAL PROTEIN DPP4
100. HYPOXANTHINE-GUANINE PHOSPHORIBOSYLTRANSFERASE
101. IMP-1 METALLO BETA-LACTAMASE
102. INDOLE-3-GLYCEROL PHOSPHATE SYNTHASE
103. INDUCIBLE NITRIC OXIDE SYNTHASE
104. INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE
105. INOSINE-ADENOSINE-GUANOSINE-PREFERRING NUCLEOSIDE HYDROLASE
106. ISOCITRATE DEHYDROGENASE
107. ISOPENTENYL-DIPHOSPHATE DELTA-ISOMERASE
108. LACTOYLGLUTATHIONE LYASE
109. LIVER GLYCOGEN PHOSPHORYLASE
110. MATRILYSIN
111. MATRIPTASE
112. METHIONINE AMINOPEPTIDASE 2
113. METHIONYL-TRNA SYNTHETASE
114. METHYLGLYOXAL SYNTHASE

115. MITOGEN-ACTIVATED PROTEIN KINASE 10

116. MITOGEN-ACTIVATED PROTEIN KINASE 14

117. MMP-13

118. MTA/SAH NUCLEOSIDASE

119. NAD-DEPENDENT FORMATE DEHYDROGENASE

120. NADH OXIDASE

121. NADPH-FLAVIN OXIDOREDUCTASE

122. NEPRILYSIN

123. NEURAMINIDASE

124. NEUTROPHIL COLLAGENASE

125. NICOTINAMIDE PHOSPHORIBOSYLTRANSFERASE

126. NITRIC-OXIDE SYNTHASE, BRAIN

127. NRH DEHYDROGENASE [QUINONE] 2

128. ORNITHINE TRANSCARBAMOYLASE

129. OROTIDINE 5'-PHOSPHATE DECARBOXYLASE

130. PEPTIDE DEFORMYLASE 1

131. PEPTIDYL-PROLYL CIS-TRANS ISOMERASE A

132. PHENAZINE BIOSYNTHESIS PROTEIN PHZD

133. PHENOL 2-HYDROXYLASE COMPONENT B

134. PHOSPHOLIPASE A2

135. PHOSPHOLIPASE C DELTA-1

136. POLY (ADP-RIBOSE) POLYMERASE

137. POLYAMINE OXIDASE

138. POLYPEPTIDE DEFORMYLASE

139. PROTEIN FARNESYLTRANSFERASE

140. PROLINE RACEMASE

141. PROTEIN KINASE B

142. PROTEIN KINASE C, THETA TYPE

143. PROTEIN KINASE CK2, ALPHA SUBUNIT

144. PROTEINASE *A (COMPONENT OF THE EXTRACELLULAR FILTRATE PRONASE)

145. PROTEIN-TYROSINE PHOSPHATASE, NON-RECEPTOR TYPE 1

146. PROTO-ONCOGENE SERINE/THREONINE-PROTEIN KINASE PIM-1

147. PROTO-ONCOGENE TYROSINE-PROTEIN KINASE ABL

148. PROTO-ONCOGENE TYROSINE-PROTEIN KINASE SRC KINASE DOMAIN

149. PURH (BIFUNCTIONAL PURINE BIOSYNTHESIS PROTEIN)

150. PURINE NUCLEOSIDE PHOSPHORYLASE

151. PYRUVATE DEHYDROGENASE E1 COMPONENT

152. RAP1A

153. RAS-RELATED PROTEIN RAL-A

154. RENIN

155. RESPIRATORY NITRATE REDUCTASE 1

156. RETINOL DEHYDRATASE

157. REVERSE TRANSCRIPTASE P66 SUBUNIT

158. RIBONUCLEASE, PANCREATIC

159. RIBULOSE BISPHOSPHATE CARBOXYLASE/OXYGENASE

160. ROUS SARCOMA VIRUS PROTEASE

161. SACCHAROPEPSIN

162. SALICYLIC ACID-BINDING PROTEIN 2

163. SARCOPLASMIC/ENDOPLASMIC RETICULUM CALCIUM ATPASE 1

164. SCYTALONE DEHYDRATASE

165. SECRETED ASPARTIC PROTEINASE

166. SERINE/THREONINE PROTEIN PHOSPHATASE 1 GAMMA (CATALYTIC SUBUNIT)

167. SERINE THREONINE-PROTEIN KINASE 6

168. SERINE/THREONINE PROTEIN PHOSPHATASE PP1-GAMMA

169. SERINE/THREONINE-PROTEIN KINASE CHK1

170. SIV PROTEASE

171. SPERMIDINE SYNTHASE

172. STROMELYSIN-1

173. TGF-BETA RECEPTOR TYPE I

174. THERMOLYSIN

175. THIOESTERASE

176. THROMBIN

177. THYMIDINE PHOSPHORYLASE

178. THYMIDYLATE KINASE

179. THYMIDYLATE SYNTHASE

180. TRIHYDROXYNAPHTHALENE REDUCTASE

181. TRNA (GUANINE-N(1)-)-METHYLTRANSFERASE

182. TRYPSIN

183. TRYPTOPHAN SYNTHASE

184. TYPE 1 17 BETA-HYDROXYSTEROID DEHYDROGENASE

185. TYROSINE-PROTEIN KINASE ITK/TSK

186. TYROSINE-PROTEIN KINASE JAK2

187. TYROSINE-PROTEIN KINASE ZAP-70

188. TYROSYL-TRNA SYNTHETASE

189. UBIQUINOL-CYTOCHROME-C REDUCTASE COMPLEX CORE, MITOCHONDRIAL

190. UBIQUITIN-PROTEIN LIGASE E3 MDM2

191. URACIL-DNA GLYCOSYLASE

192. URIDINE PHOSPHORYLASE, PUTATIVE

193. UROKINASE-TYPE PLASMINOGEN ACTIVATOR

194. WEE1-LIKE PROTEIN KINASE

195. XYLOSE ISOMERASE (GLUCOSE ISOMERASE)

HIGH-AFFINITY COMPLEXES OF NON-ENZYMES ARE COMPRISED OF:

1. ACETYLCHOLINE-BINDING PROTEIN
2. ANDROGEN RECEPTOR LIGAND BINDING DOMAIN
3. ARABINOSE BINDING PROTEIN
4. ARTIFICIAL NUCLEOTIDE BINDING PROTEIN (ANBP)
5. AUXIN-BINDING PROTEIN 1
6. AVIDIN-RELATED PROTEIN AVR4
7. BACULOVIRAL IAP REPEAT-CONTAINING PROTEIN 7 ML-IAP
8. CATION-INDEPENDENT MANNOSE 6-PHOSPHATE RECEPTOR
9. CELLULAR RETINOL BINDING PROTEIN II
10. CIRCULARLY PERMUTED CORE-STREPTAVIDIN E51/A46
11. CRABP-II
12. C-TERMINAL BINDING PROTEIN 3 (CTBP/BARS: A DUAL-FUNCTION PROTEIN INVOLVED IN TRANSCRIPTION COREPRESSION AND GOLGI MEMBRANE FISSION)

13. D-*GALACTOSE/D-*GLUCOSE BINDING PROTEIN

14. D-RIBOSE-BINDING PROTEIN

15. DIGA16

16. DODECIN

17. DUAL ADAPTOR OF PHOSPHOTYROSINE AND 3-PHOSPHOINOSITIDES

18. ESTROGEN RECEPTOR ALPHA

19. ESTROGEN RECEPTOR BETA

20. EUKARYOTIC TRANSLATION INITIATION FACTOR 4E

21. FATTY ACID-BINDING PROTEIN, BRAIN

22. FEMALE-SPECIFIC HISTAMINE BINDING PROTEIN 2

23. FIMH PROTEIN

24. FMN-BINDING PROTEIN

25. GLUTAMATE RECEPTOR 2

26. GLUTAMATE RECEPTOR, IONOTROPIC KAINATE 1

27. GLUTAMATE RECEPTOR, IONOTROPIC KAINATE 2

28. GLYCOGEN PHOSPHORYLASE (ALLOSTERIC BINDING SITE)

29. HISTIDINE-BINDING PROTEIN COMPLEXED WITH L-HISTIDINE

30. HIV-1 REVERSE TRANSCRIPTASE (NON-NUCLEOSIDE INHIBITORS)

31. HPV11 REGULATORY PROTEIN E2

32. HUMAN NUCLEAR CAP-BINDING-COMPLEX

33. IMPORTIN ALPHA-2 SUBUNIT
34. INOSITOL 1,4,5-TRISPHOSPHATE RECEPTOR TYPE 1
35. INTEGRIN ALPHA-L (LFA-1)
36. KINESIN-LIKE PROTEIN KIF11 KINESIN-MOTOR DOMAIN
37. KINESIN-RELATED MOTOR PROTEIN EG5
38. L-*ARABINOSE-BINDING PROTEIN
39. LEUCINE-SPECIFIC BINDING PROTEIN
40. LYSINE, ARGININE, ORNITHINE-BINDING PROTEIN (AMINO ACID TRANSPORT)
41. MALTOSE-BINDING PERIPLASMIC PROTEIN
42. NEUTROPHIL GELATINASE-ASSOCIATED LIPOCALIN
43. NUCLEAR RECEPTOR (STEROIDOGENIC FACTOR-1 LIGAND BINDING DOMAIN)
44. NUCLEAR RECEPTOR ROR-BETA
45. OSMOPROTECTION PROTEIN (PROX)
46. PERIPLASMIC OLIGO-PEPTIDE BINDING PROTEIN
47. PEROXISOME PROLIFERATOR ACTIVATED RECEPTOR GAMMA LIGAND BINDING DOMAIN
48. PEROXISOMAL TARGETING SIGNAL 1 RECEPTOR
49. PROGESTERONE RECEPTOR

50. PUTATIVE AMINO-ACID TRANSPORTER PERIPLASMIC SOLUTE-BINDING PROTEIN
51. RAB PROTEINS GERANYLGERANYLTRANSFERASE COMPONENT A 1
52. RETINOBLASTOMA PROTEIN
53. RETINOIC ACID RECEPTOR BETA
54. RETINOIC ACID RECEPTOR RXR-ALPHA
55. RETINOL BINDING PROTEIN
56. SEX HORMONE-BINDING GLOBULIN
57. TETRACYCLINE REPRESSOR
58. THYROID HORMONE RECEPTOR BETA-1
59. TRANSCRIPTION ELONGATION PROTEIN NUSA
60. TRANSTHYRETIN
61. TYROSINE-PROTEIN KINASE BTK
62. TYROSINE-PROTEIN KINASE TRANSFORMING PROTEIN SRC
63. VITAMIN D NUCLEAR RECEPTOR
64. VITAMIN D3 RECEPTOR
65. ANTIBODIES
    (a) 28B4 FAB (CATALYTIC)
    (b) 29G11 FAB
    (c) 4-4-20 (IG*G2A=KAPPA=) FAB FRAGMENT

(d) DIELS ALDER CATALYTIC ANTIBODY FAB

(e) ANTI-TESTOSTERONE FAB

(f) BLUE FLUORESCENT ANTIBODY (19G2)

(g) CATALYTIC ANTIBODY FAB 15A9

(h) CATALYTIC ANTIBODY FAB 34E4

(i) CHIMERIC 48G7 FAB

(j) ANTI-MORPINE FAB 9B1

(k) HLA CLASS I HISTOCOMPATIBILITY ANTIGEN WITH FAB

(l) CATALYTIC IG ANTIBODY D2.3

(m) IG KAPPA-CHAIN

(n) IMMUNOGLOBULIN

(o) IMMUNOGLOBULIN E

(p) CATALYTIC IMMUNOGLOBULIN MS6-164

(q) MONOCLONAL ANTIBODY FV4155

(r) TAB2

(s) CHA255 IMMUNOGLOBULIN

(t) IGG2B (KAPPA) FAB

(u) CATALYTIC IMMUNOGLOBULIN 6D9

APPENDIX B

MDM2 Dynamics

B.1 Introduction

The p53 tumor suppressor, also known as the guardian of the genome, is vital in cell cycle regulation, DNA repair, and apoptosis (63; 64; 65). Mutations in p53 are seen in approximately half of all human cancers (66). Where p53 is in wild-type form, it is inhibited by over-expression(67; 68) or amplification(69) of murine double minute 2 oncoprotein (MDM2; also referred to as HDM2 in human). Reactivation of p53 through inhibition of the p53-MDM2 interaction has been shown to be a novel approach for initiating or enhancing cancer cell death (70; 71). A better understanding of MDM2 dynamics is important for the design of more selective and potent inhibitors of the MDM2-p53 interaction.

A crystal structure containing residues 25 to 109 of MDM2, and residues 17 to 29 of p53, was solved in 1999 (1YCR) (72). This showed two approximately similar sub-domains, which come together to form a binding cleft for p53. Three side-chains of p53 (Phe19, Trp23, and Leu26) fill the relatively deep hydrophobic pocket. This crystal structure has been the basis of several dynamics studies (73; 74; 75; 76). In all

cases, the authors compared the MDM2-p53 complex to that of *apo*-MDM2, which was generated by removing the peptide from the structure prior to the dynamics simulation.

Barrett *et al.* utilized *CONCOORD* (77), a non-Newtonian method of ensemble generation to examine protein motion (73). They found that the principle mode of *apo*-MDM2 was a bilobal flexing, or breathing, of the protein; this motion was greatly reduced in the p53 bound complex. Previous work in our lab has utilized MD simulations to develop receptor based pharmacophore models. The models were used to identify five small-molecule inhibitors of the MDM2-p53 interaction (78; 76). Espinoza-Fonseca and Trujillo-Ferrara presented two 35-ns molecular dynamics (MD) simulations; again they demonstrated that the *apo*-MDM2 had a highly flexible and narrow cleft (75). Conversely, when the p53 peptide was bound, the cleft was more stable and wider. They also reported important side-chain motions in residues Leu57, Tyr67, His96, and Tyr100 which were present in *apo* MDM2 but not MDM2-p53, and suggested that these motions are involved in the molecular recognition of p53 and other ligands (75). A recent molecular study of the *X. laevis* by Espinoza-Fonseca and Garcia-Machorro has indicated that aromatic-aromatic interactions are involved in the interaction of p53 and MDM2 (242).

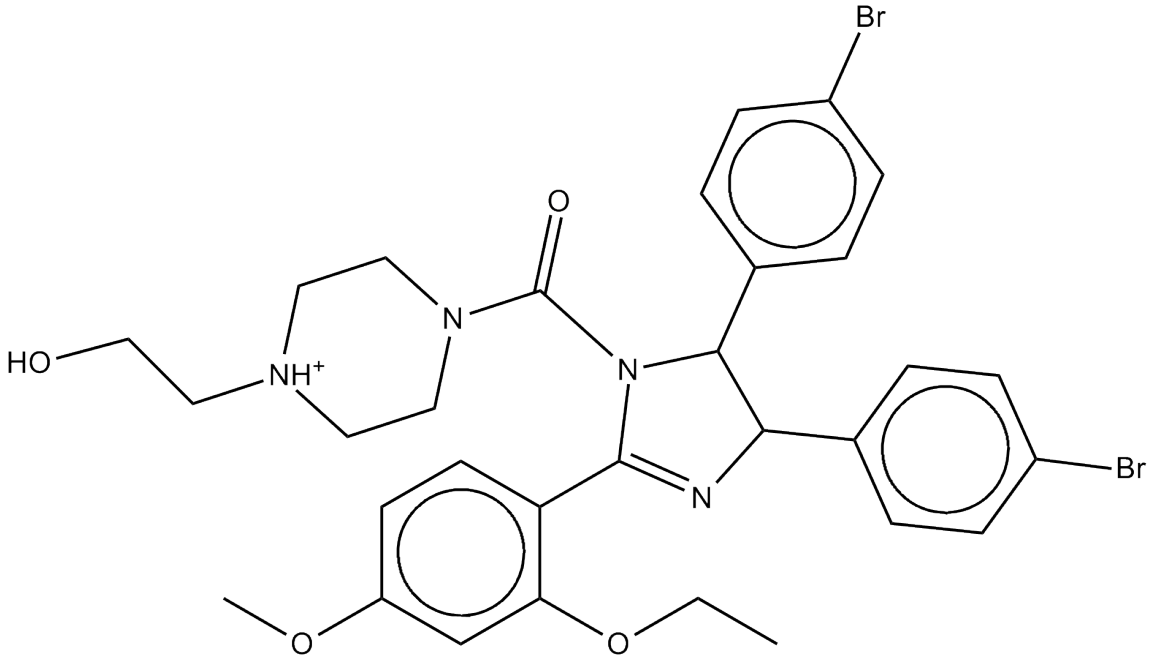
The deep, well-defined binding cleft shown in the crystal structure of MDM2-p53, suggested that the MDM2 cleft would be a suitable target for small molecule inhibitors (243). To date, several small molecule inhibitors of the MDM2-p53 interaction have been reported and the subject of recent reviews (79; 80; 81)). Through structure-based design, a nano-molar inhibitor of MDM2 (K_i = 3 nM) has been discovered (244). The crystal structure of MDM2 has been solved with both a member of the nutlin class (1RV1)(82) and a 1,4,-benzodiazepine-2,5-diones (1T4E) (83). A total of ten structures of MDM2 are found in the Protein Data Bank (115), all with a small molecule bound, and all missing the N-terminal "lid" residues.

The sequence of MDM2 residues 16-24 is highly conserved in mammals (84). NMR studies show that these residues form a lid which stabilizes MDM2 in the absence of p53 (84; 86; 85). When the lid is closed, it shields the hydrophobic binding cleft of MDM2. Ile19 occupies the same space as Pro27 of the bound p53 peptide, and makes interactions with His96, Arg97, and Tyr100 (85). The lid unfortunately is not well-defined in the *apo*-NMR structure, 1Z1M, but this was the first and only structure including it when we began this study (85). NMR studies on MDM2 residues 17-125 indicate the *apo*-MDM2 structure exists in two states, with the dominant state being closed, and the minor state open (245). There is also evidence that the lid is phosphorylated on Ser18 (86). When the lid is phosphorylated it binds the pocket and inhibits p53 association (86). The unphosphorylated lid is easily displaced by p53, as well as small molecule and peptidic inhibitors (84). It is thought that the two sub-domains then swing apart by 3-4 Å to deepen the binding cleft, allowing the peptide or inhibitor to completely bind (84; 86; 85). In this study, we present the preliminary results of molecular dynamics simulations of the full-length *apo* N-terminal of MDM2 and the nutlin inhibitor from the x-ray crystal structure, which has an IC_{50} of 0.14 μM (82). The structure of the nutlin is show in Figure B.1.

B.2 Methods

An NMR ensemble of the *apo* N-terminal, p53 binding region of MDM2 is available (1Z1M) (85). This ensemble consists of 24 structures and provided the starting conformations for the protein in each set of Langevin dynamics simulations. The NMR structure was from a more complete construct (residues 2-118) than that used for the available x-ray structures. However, residues 2-6 had no assignments and were assumed to be disordered (85). Therefore, the N-terminal residues 1-6, Met1, Val2, Arg3, Ser4, Arg5, and Gln6, were built in with PyMOL (141). The resulting protein represents the entire p53-binding region, including the binding site lid.

Figure B.1: Structure of nutlin inhibitor used in molecular dynamics simulation.



The crystal structure of a nutlin small molecule with MDM2 (1RV1)(82) was overlaid on each structure from the NMR ensemble. The MDM2 protein was removed from 1RV1, leaving just the nutlin in the binding cleft of the NMR structure. To attempt to recreate a binding event, the nutlin was then translated to two positions where the inhibitor RMSD values from the x-ray structure were 13.0 and 16.8 Å, but the original orientation of the ligand was preserved. From these conformations, two positions other were obtained by rotating the ligand 180 degrees on itself after translation, yielding nutlin RMSD values of 15.7 and 19.3 Å respectively.

Hydrogen atoms were placed by using the LEaP (246) module in AMBER (247). FF03(248) was used together with parameters for the nutlin were obtained using the antechamber module and GAFF(249) with AM1-BCC charges (250). Langevin dynamics were performed, with a collision frequency of 1 ps^{-1}, and a timestep of 1 fs. No nonbonded cutoff was applied. Aqueous solvation was modeled implicitly with a modified generalized Born model (251). The system was minimized, and then gradually heated to 300K. All heavy atoms were harmonically restrained, with the

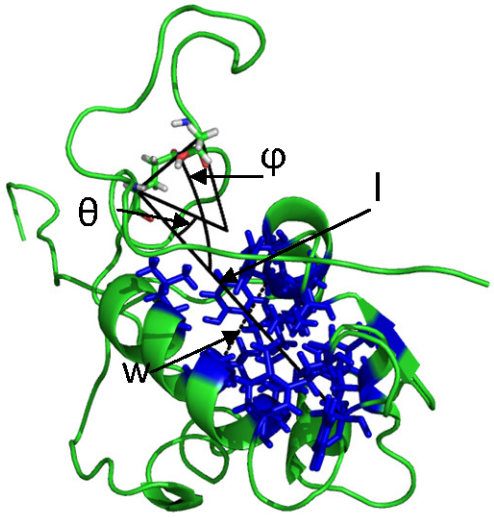
force decreasing from 2 to 0.1 kcal/mol-Å^2 in four steps, for a total of 80 ps. Then just the backbone atoms were harmonically restrained at 0.1 kcal/mol-Å^2 for 50 ps. The last step of equilibration was 100 ps of unrestrained Langevin dynamics at 300K. All production simulations were unrestrained at 300K for 1 ns.

The ptraj module in AMBER was used to generate snapshot structures every 1 ps from each simulation. These structures were overlaid using a Gaussian-weighted alignment tool (252). This superimposed the core of the protein, without the highly flexible N- and C-terminus tails skewing the alignment. The MMTSB toolset(253) was used to calculate the RMSD of the flexible residues in the binding site: Leu63, Tyr73, His96, and Tyr100. It was also used to calculate the RMSD of the nutlin ligand in each structure to the position of the ligand in the crystal structure. The center of mass of the nutlin ligand was also calculated and compared to the center of mass in the crystal structure.

The dynamics of the cleft were estimated by using NACCESS(50) to calculate the solvent accessible surface area of the residues lining the binding pocket, while ignoring the ligand (colored in blue in Figure B.2). The width (w) of the binding site is the length of the vector between the Cα of Leu63 and the Cα of Tyr100. The length (l) of the binding site is the length of the vector between the Cα of Gln24 at the base of the lid and the Cα of Tyr73 at the other end of the binding cleft (see Figure B.2).

The dynamics of the cleft with respect to the binding site are determined by two angles, the angle of the lid with respect to the cleft and the shear angle with respect to the plane of the binding cleft. The (φ) angle is formed between the lid vector formed by the Cα of Gln22 and the Cα of Thr16 or the center of mass of the lid residues (residues 16-24). For our discussion we use the center of mass of the lid residues since it is most representative of the lid. The shear of the binding site (θ) is the angle projected by the lid vector on the plane of the binding site formed by the width and length vectors. See Figure B.2 for more details. A completely closed lid will have φ

Figure B.2: This figure shows the definition of the angles (θ and φ of the lid with respect to the binding pocket. The length(l) and width(w) of the binding pocket are shown. The residues on the surface of the binding pocket are colored blue.



= 0 and $\theta = 0$. The distance from this point is then defined by d = $\sqrt{(\varphi^2 + \theta^2)}$ and will be used as a direct measure of openness.

B.3 Result and Discussion

B.3.1 Lid Dynamics

Initial comparisons of the RMSD of the ligand in the snapshots compared to the RMSD of the ligand in the x-ray structure indicate four simulations were able to reproduce the binding of the inhibitor into the binding site (RMSD< 3 Å). One additional structure reproduced the fit but only stayed in the bound conformation for a short period of time at the end of equilibration. It is left off of Figure B.3, since the inhibitor is not in the bound conformation during nearly the entirety of the production run. The dynamics of the lid in each of the simulations that reproduce the pose varies. In the simulation which obtained the ligand pose closest to the small molecule in the x-ray, the lid opens very wide, which allows the small molecule to

enter the pocket (Figure B.3A). This is the only structure to reproduce the crystal structure, but also have the lid open. The other three cases display of variety of lid dynamics. In fact, during the second trajectory the small molecule gets caught in the N-terminal residues and is brought close to the pocket by these residues (Figure B.3C). In the third simulation the lid is just open enough to allow the ligand to enter the binding site, and then closes down on the ligand and binding site. In the fourth system the lid opens during equilibration then comes closes on part of the pocket when the small molecule binds.

The lid has been shown previously to have large fluctuations upon small molecule binding. In an NMR study by Showalter *et al.* the *apo-* form of MDM2 was shown to predominantly favor a closed lid state, while the state with 13-residues of p53 bound was only found in the open lid conformation (245). Coordinates for this structure were not deposited in the PDB for comparison. However, the bound state does not appear to be open in the four simulations that represent the nutlin binding mode. In the third best simulation, that represents the bound state of the ligand, the lid has closed on the small molecule. Figure B.4 further demonstrates this since the majority of structures with a center of mass near the ligand in the crystal structure have a relatively closed lid. Additionally, all snapshots where the ligand is unbound (> 40 Å from the x-ray pose) have a lid that is open, which contradicts the NMR findings, but since there are so few snapshots that are unbound, compared to bound, it may be an artifact of the ligand in the simulation, which on occasion interacts with unstructured N-terminal residues, influencing the lid.

What also can be seen in Figure B.4 is that many of the snapshots have a small molecule within 5 Å of the crystal structure ligand, but only four simulations were able to reproduce the binding mode of the crystal structure. This indicates that more sampling may be necessary in these simulations to allow the ligand to adopt the appropriate conformation. In addition, since the lid is shown to bind in the

Figure B.3: This figure shows the RMSD between the snapshot pose and the x-ray pose of the nutlin inhibitor, and the openness of the lid in each snapshot of the four simulations that reproduced the ligand pose of the x-ray crystal structure (1RV1). Figure A is the simulation that had the smallest RMSD, and Figures B, C and D had the second, third and fourth smallest RMSD, respectively. Negative time from the simulation indicates the equilibration time period. This is included in the figure since the ligand does interact with the protein before the start of the production run.

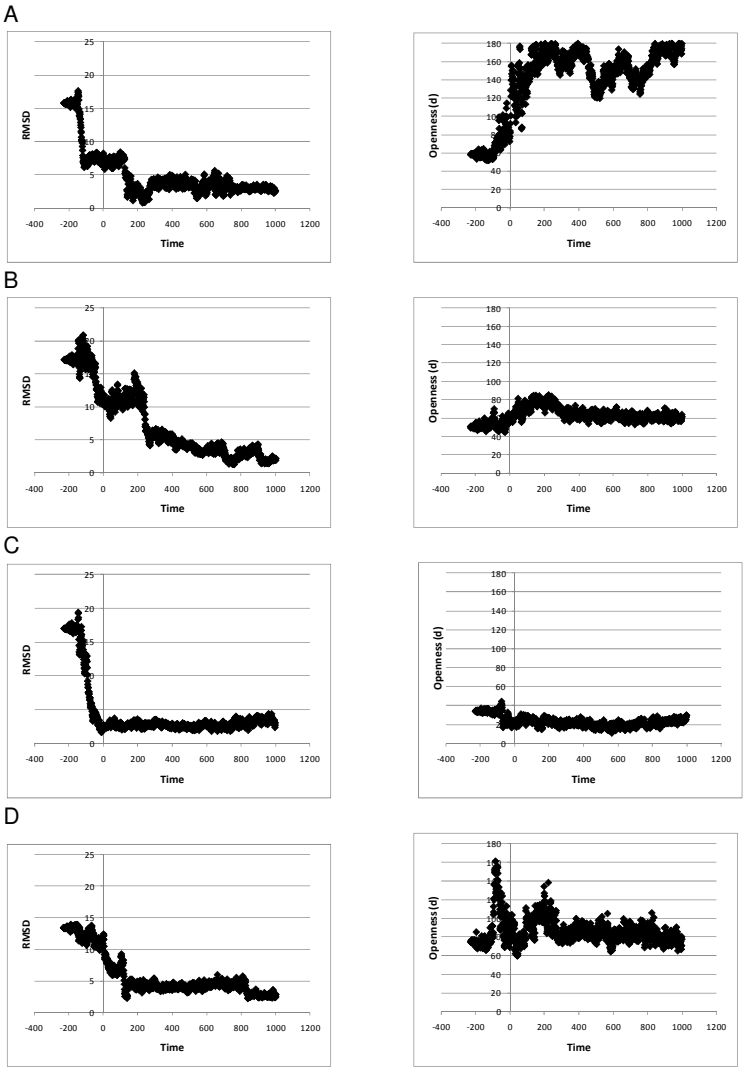
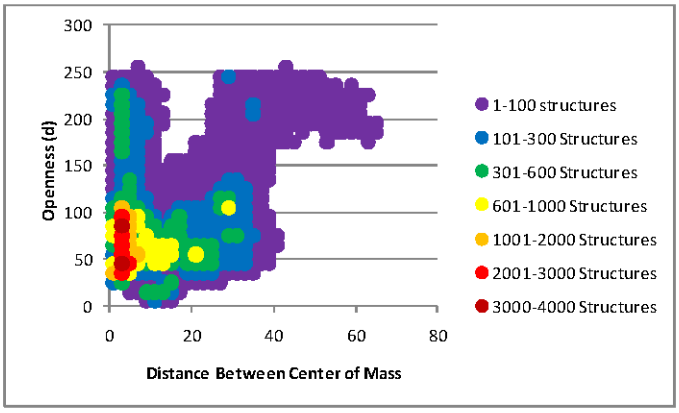


Figure B.4: This figure is a histogram of the distance between the center of mass of the inhibitor in the snapshot and the center of mass of the inhibitor in the x-ray structure (1RV1), and the openness of the pocket. The colors represent the number of structures in each bin. Bins were created for every 2 Å for the distance between the center of mass and 20 degrees for the openness of the lid.



pocket when it is phosphorylated and p53 does not bind, the frequency at which it is phosphorylated may have some impact on lid dynamics upon small-molecule binding (86).

B.3.2 Pocket Dynamics

In each of the four cases that reproduced the conformation of the ligand in the binding site, the pocket begins in an exposed conformation (˜500 Å^2). In three of the four cases the pocket decreases in exposed surface by ˜200-300 Å^2 when the ligand bound, Figure B.5. Although in the simulation that most closely reproduces the crystal structure position of the ligand, the pocket remains somewhat exposed with SASA ≈400 Å^2, which may be due to the lid being completely displaced in this simulation (see Figure B.3). The change in SASA upon binding is in agreement with previous molecular dynamics simulations by Espinoza-Fonseca and Trujillo-Ferrara (75). They have noted that the SASA of the binding pocket is ˜100 Å^2 less in the complex bound to a p53 peptide compared to the *apo*-MDM2 system. Their study

cites an experimental decrease in solvent accessible surface area of ˜200 Å^2, but do not a provide reference for the value (75). The change in surface area calculated in their study is with respect to all the residues in the structure. However, we aimed to investigate the surface area of the pocket, not the entire protein, and therefore only calculated the SASA of the residues in the binding site. Our finding is also consistent with high-temperature molecular dynamics simulations to investigate unfolding of MDM2 performed by Chen and Luo, in which MDM2, upon unfolding, tertiary contacts decrease as the number of native binding contacts of the p53 peptide decrease (74). They suggested that this is, in actuality, a folding pathway as the p53 peptide binds. Therefore, ligand binding induces folding of MDM2 (74).

Only the four simulations mentioned above of the 96 simulations have reproduced the binding site ligand conformation. Therefore, we looked at the ensemble of all snapshots across all simulations. From Figure B.6, we see that as the ligand approaches the pocket, or as the center of mass is getting closer to the bound ligand, (up until about 20 Å from the pocket) the SASA remains about the same, but once inside that range the pocket begins to become less exposed. The calculations of SASA were performed with the ligand removed from the structure, therefore the change in SASA of the residues around the pocket reflects a "closing" of the pocket. Breaking the SASA down into contributions from the individual residues is possible. Future analysis will break down the SASA of the binding pocket. This may indicate which residues are contributing most to the change in surface area, and show which residues in the binding site are most flexible and/or necessary for binding. We hypothesize that the most flexible residues would be Leu57, Tyr67, His96, and Tyr100, as they were determined to change the most upon p53 peptide binding by Espinoza-Fonseca and Trujillo-Ferrara (75). This can be confirmed by examining the root-mean-square fluctuation of these residues as the small molecule binds to the pocket. However, initial investigations of the RMSD of these four residues do not indicate significant

Figure B.5: This figure shows the RMSD between the snapshot pose and the x-ray pose of the nutlin inhibitor, and the exposed surface area of the pocket in each snapshot of the four simulations which reproduced the ligand pose of the x-ray crystal structure (1RV1). Figure A is the simulation that had the smallest RMSD, and Figures B, C and D had the second, third and fourth smallest RMSD, respectively. Negative time from the simulation indicates the equilibration time period. This is included in the figure since the ligand does interact with the protein before the start of the production run.

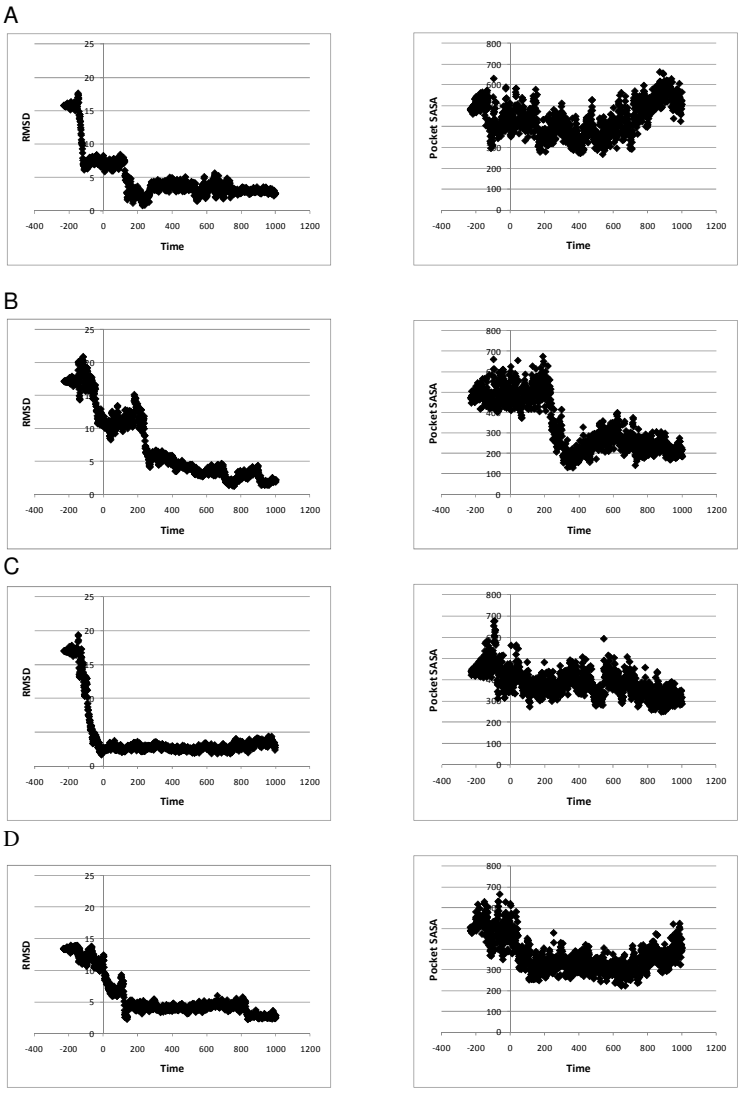
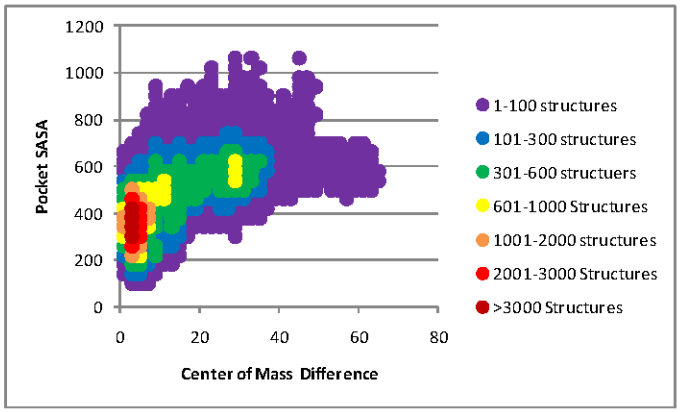


Figure B.6: This figure is a histogram of the distance between the center of mass of the inhibitor in each snapshot and the center of mass of the inhibitor in the x-ray structure (1RV1), and the SASA of the binding pocket. The colors represent the number of structures in each bin. Bins were created for every 2 Å for the distance between the center of mass and 40 Å^2 for the SASA of the pocket.



dynamics.

B.4 Conclusion

From molecular dynamics simulations of the complete, *apo*-MDM2 NMR structure we were able to investigate the dynamics of the lid and the binding pocket as a small molecule enters the binding site. We notice that the lid provides different dynamics than previously indicated when the p53 peptide bound. Here the lid appears to be in a closed state when the small molecule is bound, and is potentially open when unbound. Secondly, upon investigation of the binding pocket, it appears that as the small molecule nears the pocket, the pocket closes and becomes less exposed to solvent. Future directions of the study will continue to investigate the behavior of the residues that line the pocket. Knowledge of the residues that display a change in their SASA upon binding may provide insight into important motions in the binding of the nutlin inhibitor.

APPENDIX C

Protein Data Bank Filtering and Updating Binding MOAD

First download the PDB, filter the PDB entries through a series of perl scripts, and then hand check them. After these scripts, the PDBs are aligned against each other using BLAST to put them into protein families of high sequence similarity. The data is then ready to put up on wasabi (on bindingmoad.org).

C.1 Download PDB Files

1. Download the latest PDBs from the Protein Databank using rsync

   source moad_update_scripts/rsync_PDB.justPDB.sh

   This script is downloaded from www.rcsb.org, and modified to have the correct paths and directories on curry.

C.2 Filtering the PDB files

1. Create a list of all PDB files.

   ls pdb_directory > pdb_date.txt

2. Open "pdb_date.txt" in vi or an editor of your choice and remove the .ent extensions and save it.

   diff MOAD_[last_year]/[last_year]_all_entries_already_ran > differences_file

   grep '>' differences_file > new_PDBs_date.txt

3. Open the "new_pdbs_date.txt" file and remove the leading '>', and save it.

4. Run the MOAD filtering scripts, these will take a day or so depending on how you break it up):

5. The caution_file and the caution_covalency_file have some duplicates. Therefore you must merge the accept, caution, anc caution covalency files into one. To do this I have written merge_accept_caution_files.pl. The syntax for running this is:

6. Make a list of PDBS to check.

C.3 Scrape HTML and load data into BUDA

1. Get a list of Pubmed IDs from NCBI and RCSB.

2. Merge the two lists (giving preference to NCBI over RCSB, as they have appeared to be more recent, and more correct)

3. Get a list of the DOIs using the list of Pubmed IDs,

4. Get the HTML files using the DOIs.

5. Rename the DOI files based on Pubmed ID.

C.4 Checking the heavy atoms

1. Run the get_hets.pl script which will produce a file of all of the het groups in the particular pdb files (This may take a while depending on how many new structures there are):

2. Run the unique_hets.pl script to only make a list of the unique het groups.

3. Run the check_heavy_atoms.pl. Be sure to change the names of the output files that are listed on lines 51 and 52 of the script.

4. These two lists are checked by hand to see what is wrong and why there are too many or too few atoms, this information should also be entered on columns M (comment) and N (percentage missing or too many) on the .csv spreadsheet generated in the first section.

5. Find the differences between the new het list and the old list of het groups that have been checked. These are HETs that still need to be checked for validity.

6. Create a new .csv which has a comment regarding the new het atoms to check.

   The new hets should be in a list with a single line for a het group with no spaces.

C.5 Literature Searching

1. Log into BUDA (The web address will be given by Peter Dresslar).

2. Go through each pdbid, if it says caution in the comment section of the ligand (or the first column has a 2 or a 3 three, check to see if the ligand/structure is valid and that it is not just a crystal additive, otherwise check to see if the ligand is on the list of new HET groups. If the ligand is on the list of new HET groups make a decision if it should be rejected, cautioned, or accepted

HET group (if it contains a metal check to see if part of the HET group should be considered a ligand). If everything is valid search the paper for a binding constant (K_i, K_d, or IC_{50}). Search terms that I use to go through the .pdf file are and look in the paragraph and sentences for the actual values:

(a) Kd

(b) Ka

(c) Ki

(d) ic50

(e) binding constant

(f) association constant

(g) dissociation constant

(h) inhibition constant

(i) association

(j) dissociation

(k) binding affinity

(l) affinity

(m) equilibrium constant

(n) free energy of binding

(o) binds with millimolar affinity

(p) binds with micromolar affinity

(q) binds with nanomolar affinity

(r) Km

3. Once you find a binding constant, check to see what ligand and structure it is associated with. The value must be for the same organism and form of the

protein (mutant or wild-type). If multiple binding constants exist choose the one that was determined at experimental conditions that most closely resembles the crystal conditions, with preference of Kd over Ki over IC50.

4. Mark the entry as "COMPLETED"

C.6 Export Entries

1. Export the entries and download the "bindings" and "messages" files.

2. Run the create_csv_for_binning.pl script with the csv's fom the export.

3. Remove any structures that have been rejected in BUDA.

4. This list also has entries from previous years, therefore you need to use the list of structure we had done in previous years to take them off the list. This only should be the structures from the new update. This will then be used for the merge.py script after binning.

C.7 Pre-Binning

1. Get the list of PDBs in current moad (with obsoletes remove) from directory.

2. Obtain the old csv from current MySQL database.

3. Obtain the list of obsolete entries from the Protein Data Bank (rcsb.org) and remove any obsoleted entries from the old csv file.

4. Make a list of the old (without the obsolete) and new entries in MOAD.

C.8 Binning

1. Run the script to get the ec numbers and fasta information: Running this script produces three files: all.txt, ec.txt, and pdbchains.fasta - these are used for the binning script.

2. Run getNonRedundantEntries.py script.

C.9 Merge Bins

1. The binning makes a directory called CurrentRun in the directory the program was ran from. The new tab delimited file is CurrentRun/blastlist/scale2/90ec-subbins.fancy.out This file needs to be copied to 90ec-subbins.fancy.edited.out, and duplicate bins need to be removed. A list of the duplicate bins is located at CurrentRun/blastlist/scale2/90ec-subbins.redundancy-warnings.out. The file '2008_hand_editing.txt' has the list of bins and why they were deleted in the past and can guide you through the hand editing process. When editing the 90ec-subbins.fancy.out make sure there is a tab after each pdbid. The merge.py script will choke otherwise.

2. Run the merge.py script to combine the old and new data. Open merge.py and change the variables on lines 62-65 to match the appropriate names. Then from the command line type 'python /users/dicksmit/src/PLD/merge.py

3. Choosing Bin leaders: Open the new output file in Excel. For each of the subbins if the subbin has an N in column F, a new leader must be chosen. The rules for selection of bin leader are as follows:

   (a) Highest affinity

   (b) Ligand present over only biological cofactor (ATP, GTP, NAD, FAD, PLP, etc.)

(c) Best Resolution

(d) Wild-type over mutant protein

(e) Human over non-human protein

(f) Deposition Date

(g) Other criteria (Comments in papers, R-value, chemical intuition)

C.10 Getting Protein Information

1. Create a file with all the pdb information. To create this file run:

   You will need this file to help classify the proteins that dont have ec numbers.

C.10.1 Classifying the enzymes that do not have EC numbers

2. Run the script to get the information for proteins that have already been classified.

3. Use the new entries output from step 1 and the output file from "Getting Protein Information" to organize the information for each of the new proteins. The output file is the list of entries that do not need to be checked.

4. With this output run the script to identify proteins. This script puts each protein into one or more of twelve categories. It also produces a summary output of which category each protein falls into (it is tab delimited

   (pdb_category.id_prot_(date)). You may want to change the output file names in the script (identify_proteins.pl) on lines 24-36 to reflect the new date. To run the script

5. Open this new file in excel and sort the list by column A, this will tell you which structures were identified in more than one group, read the paper and

decide which class the protein should be in, change the name in column C to reflect your choice (binding, signal_hormone, enzyme, mobile, transport, immune, structural, transcript_translate, toxin_viral, folding, cell_cycle, other) Then sort the list by column C, and take a look at the ones that were categorized as enzyme, or other and see if they could be given EC number (1.-.-.-: oxidase, 2.-.-.-: transferase, 3.-.-.-: hydrolase, 4.-.-.-: lyase, 5.-.-.-: isomerase, 6.-.-.-: ligase) or placed in one of the other 11 classes (other than other). Use the the list of new entries from step 2 to help you decide since that has the proteins already subbin (you should have proteins in the same subbin have the same classification). Change the unmatched classification in the new entry file from step 2 to reflect the new subbins classification.

6. Concatenate the modified file of the new entries with the entries that already were classified, to create the new classification file.

7. Make the final file into a comma separated file by using your favorite text editor. You must also make sure the ec numbers are in the format 1.-.-.-, 2.-.-.- etc.

C.11 Processing the Biounit Files

C.11.1 Make Biounits

1. Download the biounit files from the PDB and copy the biounits for the new structures in MOAD to a new directory.

2. Remove atoms that are greater than 10 Åfrom any protein atoms.

3. Determine which biounit files contain multiple models; these files will need the waters rotated. Move thes fixed files corresponding to those structures into a separate directory.

4. Go into this new directory and run the rotate_waters.py script.

5. Make a new directory to put these new files in, they are currently placed in the same directory as the files with the 10 angstroms removed.

6. Move the output files from step 5 to the new directory.

7. Run the script to get the header information from the fixed file. The script that removes the waters renames some of the chains so you must tell it where the fixed files are and not the original biounit file.

8. Copy those files with the header to a new directory.

9. Pymol chops the ends of the lines off the pdb format so you must add spaces to the end of the file to have a total of 80 characters.

10. Remove any waters that ended up more than 10 Åfrom the protein just as was done in step 3.

11. Rename the final outputs [pdb_id].bio[#].

C.12 Generate Multi-Part SMILES

1. Figure out which structures have multi-part ligands.

2. Copy biounit files of these structures to a new directory.

3. Make ligand files for these structures This creates a new directory called ligands within the biounit directory

4. Run the perl script which runs the svl script to generate the SMILES string. First you must change the file for the .mdb file on line 44 of the save_db.svl script.

5. Copy the .mdb file over to a windows machine and open it in MOE. Then save it as a '.csv' file.

6. Use a text editor to only keep the 'ligand name, SMILES'. Remove the pdbid, chain name, residue number and other information from the first column (before the first comma) leaving only the ligand name. After the first comma delete everything before the SMILES string.

7. Concatenate this new file with the old file

C.13 Run Gocav on new biounits

1. Run the script to run gocav over all the structures in a directory. You may do this on as many nodes as you would like to split up the list., be sure you run it from the directory you would like to put the output (script on curry)

2. Run script to rename the files (script on curry).

3. Make a new directory to put divided entries in, and make two subdirectories, nowater and water.

4. Run script to divide the files into a new directory (script on curry)

5. Run script to copy the files to the format for the viewer (script on curry).

6. Choose only one ligand for jar files.

7. Create linking shell script for jar files

8. Make directory for jar files

9. Change name of directories in script to reflect the desired xyz and jar directories.

10. move the jar files to /users/moaddata/jar

C.14 Create the extracted references.csv and authors.csv

1. Change line 30 in extract_reference.pl (found in

   /home/moad_2008_update/moad_update_scripts on basil) to reflect the list of all structures in binding moad.

   (variable $source_file)

2. Change line 35 in extract_reference.pl to reflect where the pdb lives (variable $pdb_path)

3. Run the extract_reference script.pl

4. Change line 11 in create_authors.pl (found in same directory as

   extract_reference.pl) to reflect name of output file from previous step, variable $file), and change line 12 to reflect where you would like the output to go ($output).

5. Run the create_authors.pl

C.15 Create NEW Database and Load Data

1. Creating a mysql dump of the old data.

2. Create the new mysql database moad_devel with password s3crtP2ss and user moad. Also create a database jboss with password jbossmoad, and user jb.

3. The files you will need to run the loading scripts need to be in the base directory where the run_marathon_scripts directory is located. These are:

   (a) The final csv.

   (b) The final list of rejected structure.

   (c) The list of protein classifications.

(d) The list of multi-part smiles.

(e) The extracted references files and the authors.csv files from the previous steps.

4. Change the name of any input files in the directory run_marathon_scripts to reflect where the new files are located.

5. Run run_marathon.sh

C.16 Zip Biounits

1. You must first generate a list of the biounit files that have valid ligands. The 'old list' is already located in /data/sandbox/2008_VALID_BIOUNIT_FILES.

2. Run generate_biounit_zips.pl

3. To genereate the new_for_xxxx.zip, run

   generate_new_biounit_zips.pl

4. Remove all files except new_for_xxxx.zip and the HiQ set of zips from

   /users/moaddata/biou

5. Copy the new_for_xxxx.zip and zip directory to /users/moaddata/biou

C.17 Generate CSV files for downloadinig

1. First make the directories class/ family/ total/ and individual/ in whatever directory you would like to work in.

2. Change the variable $license_file in

   generate_csv_from_mysql.pl to reflect the location of the license file. The file is currently on basil in /data/sandbox/BindingMOAD_license.txt

3. Run the script.

4. Generate a '.csv' of only the new entries

5. First change line 20 to reflect where the license file is, and change line 218 on generate_new_entry_csv_from_mysql.pl to reflect where and what you would like the new file to placed and named. Then run the script.

6. Remove all files from the /users/moaddata/csv except the new_for_xxxx.csv files

7. Copy the files from in class family individual and total to /users/moaddata/csv. Also copy the file created from step three to

   /users/moaddata/csv/new_for_xxxx.csv

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Stiffin, R. M., Sullivan, S. M., Carlson, G. M., and Holyoak, T. (2008) Differential inhibition of cytosolic pepck by substrate analogues. kinetic and structural characterization of inhibitor recognition. *Biochemistry 47*, 2099–2109.

[2] Ash, D. E., Emig, F. A., Chowdhury, S. A., Satoh, Y., and Schramm, V. L. (1990) Mammalian and avian liver phosphoenolpyruvate carboxykinase. alternate substrates and inhibition by analogues of oxaloacetate. *J Biol Chem 265*, 7377–7384.

[3] Fischer, E. (1894) Einfluss der configuration auf die wirkung der enzyme. *Ber. Dtsch. Chem. Ges. 27*, 2985–2993.

[4] Gohlke, H. and Klebe, G. (2002) Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew Chem Int Ed Engl 41*, 2644–2676.

[5] Kunz, H. (2002) Emil fischer–unequalled classicist, master of organic chemistry research, and inspired trailblazer of biological chemistry. *Angew Chem Int Ed Engl 41*, 4439–4451.

[6] Lipscomb, W. N. (1994) Linus pauling 1901 - 1994. *Structure 2*, 991–991.

[7] Pauling, L. (1948) Nature of forces between large molecules of biological interest. *Nature 161*, 707–709.

[8] Koshland, D. E., Ray, W. J., and Erwin, M. J. (1958) Protein structure and enzyme action. *Fed Proc 17*, 1145–1150.

[9] Carlson, H. A. (2002) Protein flexibility and drug design: how to hit a moving target. *Curr Opin Chem Biol 6*, 447–452.

[10] Carlson, H. A. and McCammon, J. A. (2000) Accommodating protein flexibility in computational drug design. *Mol Pharmacol 57*, 213–218.

[11] Luque, I. and Freire, E. (2000) Structural stability of binding sites: consequences for binding affinity and allosteric effects. *Proteins Suppl 4*, 63–71.

[12] Ma, B., Shatsky, M., Wolfson, H. J., and Nussinov, R. (2002) Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci 11*, 184–97.

[13] Teague, S. J. (2003) Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov 2*, 527–41.

[14] Gilson, M. K. and Zhou, H. X. (2007) Calculation of protein-ligand binding affinities. *Annu Rev Biophys Biomol Struct 36*, 21–42.

[15] Aqvist, J., Medina, C., and Samuelsson, J. E. (1994) A new method for predicting binding affinity in computer-aided drug design. *Protein Eng 7*, 385–91.

[16] Xu, J., Deng, Q., Chen, J., Houk, K. N., Bartek, J., Hilvert, D., and Wilson, I. A. (1999) Evolution of shape complementarity and catalytic efficiency from a primordial antibody template. *Science 286*, 2345–2348.

[17] Liang, J., Edelsbrunner, H., and Woodward, C. (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci. 7*, 1884–1897.

[18] Ji, T. H., Grossmann, M., and Ji, I. (1998) G protein-coupled receptors. i. diversity of receptor-ligand interactions. *J Biol Chem 273*, 17299–17302.

[19] Lafont, V., Armstrong, A. A., Ohtaka, H., Kiso, Y., Amzel, L. M., and Freire, E. (2007) Compensating enthalpic and entropic changes hinder binding affinity optimization. *Chem Biol Drug Des 69*, 413–422.

[20] Fonseca, T., Ladanyi, B. M., and Hynes, J. T. (1992) Solvation free energies and solvent force constants. *J Phys Chem 96*, 4085–4093.

[21] Miyamoto, S. and Kollman, P. A. (1993) What determines the strength of noncovalent association of ligands to proteins in aqueous solution? *Proc Natl Acad Sci U S A 90*, 8402–6.

[22] DeChancie, J. and Houk, K. N. (2007) The origins of femtomolar protein-ligand binding: hydrogen-bond cooperativity and desolvation energetics in the biotin-(strept)avidin binding site. *J Am Chem Soc 129*, 5419–29.

[23] Hyre, D. E., Le Trong, I., Merritt, E. A., Eccleston, J. F., Green, N. M., Stenkamp, R. E., and Stayton, P. S. (2006) Cooperative hydrogen bond interactions in the streptavidin-biotin system. *Protein Sci 15*, 459–67.

[24] Abad-Zapatero, C. and Metz, J. T. (2005) Ligand efficiency indices as guideposts for drug discovery. *Drug Discov Today 10*, 464–9.

[25] Hopkins, A. L., Groom, C. R., and Alex, A. (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discov Today 9*, 430–431.

[26] Rees, D. C., Congreve, M., Murray, C. W., and Carr, R. (2004) Fragment-based lead discovery. *Nat Rev Drug Discov 3*, 660–672.

[27] Kuntz, I. D., Chen, K., Sharp, K. A., and Kollman, P. A. (1999) The maximal affinity of ligands. *Proc Natl Acad Sci U S A 96*, 9997–10002.

[28] Chong, L. T., Dempster, S. E., Hendsch, Z. S., Lee, L. P., and Tidor, B. (1998) Computation of electrostatic complements to proteins: a case of charge stabilized binding. *Protein Sci 7*, 206–10.

[29] Kangas, E. and Tidor, B. (1999) Charge optimization leads to favorable electrostatic binding free energy. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 59*, 5958–61.

[30] Lee, L. P. and Tidor, B. (2001) Optimization of binding electrostatics: charge complementarity in the barnase-barstar protein complex. *Protein Sci 10*, 362–77.

[31] Kangas, E. and Tidor, B. (2001) Electrostatic complementarity at ligand binding sites: Application to chorismate mutase. *JOURNAL OF PHYSICAL CHEMISTRY B 105*, 880–888.

[32] Fuhrmann, C. N., Daugherty, M. D., and Agard, D. A. (2006) Subangstrom crystallography reveals that short ionic hydrogen bonds, and not a his-asp low-barrier hydrogen bond, stabilize the transition state in serine protease catalysis. *J Am Chem Soc 128*, 9086–102.

[33] Schott, B., Iversen, B. B., Madsen, G. K. H., Larsen, F. K., and Bruce, T. C. (1998) On the electronic structure of low-barrier hydrogn bonds in enzymatic reactions. *Proc Natl Acad Sci U S A 95*, 12799–12802.

[34] Smith, A. J., Zhang, X., Leach, A. G., and Houk, K. N. (2009) Beyond picomolar affinities: quantitative aspects of noncovalent and covalent binding of drugs to proteins. *J Med Chem 52*, 225–33.

[35] Gitlin, I., Carbeck, J. D., and Whitesides, G. M. (2006) Why are proteins charged? networks of charge-charge interactions in proteins measured by charge ladders and capillary electrophoresis. *Angew Chem Int Ed Engl 45*, 3022–60.

[36] Frank, H. S. and Evans, M. W. (1945) Free volume and entropy in condensed systems. iii. entropy in binary liquid mixtures; partial entropy in dilute solutions; strucure and thermodynamics in aqueous electrolytes. *Journal of Chemical Physics 13*, 507–532.

[37] Sharp, K. A., Nicholls, A., Fine, R. F., and Honig, B. (1991) Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science 252*, 106–109.

[38] Olsson, T. S., Williams, M. A., Pitt, W. R., and Ladbury, J. E. (2008) The thermodynamics of protein-ligand interaction and solvation: insights for ligand design. *J Mol Biol 384*, 1002–17.

[39] Chervenak, M. C. and Toone, E. J. (1994) A direct measure of the contribution of solvent reorganization to the enthalpy of ligand binding. *Journal of the American Chemical Society 116*, 10533–10539.

[40] Doig, A. J. and Sternberg, M. J. (1995) Side-chain conformational entropy in protein folding. *Protein Sci 4*, 2247–51.

[41] Pickett, S. D. and Sternberg, M. J. (1993) Empirical scale of side-chain conformational entropy in protein folding. *J Mol Biol 231*, 825–39.

[42] Yu, L., Zhu, C. X., Tse-Dinh, Y. C., and Fesik, S. W. (1996) Backbone dynamics of the c-terminal domain of escherichia coli topoisomerase i in the absence and presence of single-stranded dna. *Biochemistry 35*, 9661–6.

[43] Zidek, L., Novotny, M. V., and Stone, M. J. (1999) Increased protein backbone conformational entropy upon hydrophobic ligand binding. *Nat Struct Biol 6*, 1118–21.

[44] Yang, C.-Y., Wang, R., and Wang, S. (2005) A systematic analysis of the effect of small-molecule binding on protein flexibility of the ligand-binding sites. *J Med Chem 48*, 5648–5650.

[45] Yang, C. Y., Sun, H., Chen, J., Nikolovska-Coleska, Z., and Wang, S. (2009) Importance of ligand reorganization free energy in protein-ligand binding-affinity prediction. *J Am Chem Soc 131*, 13709–21.

[46] Raag, R. and Poulos, T. L. (1991) Crystal structures of cytochrome p-450cam complexed with camphane, thiocamphor, and adamantane: factors controlling p-450 substrate hydroxylation. *Biochemistry 30*, 2674–84.

[47] Levitt, D. G. and Banaszak, L. J. (1992) Pocket: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J.Mol.Graphics 10*, 229–234.

[48] Laskowski, R. A. (1995) Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J.Mol.Graphics 13*, 323–330.

[49] Brady, G. P. and Stouten, P. F. (2000) Fast prediction and visualization of protein binding pockets with pass. *J Comput Aided Mol Des 14*, 383–401.

[50] Hubbard, S. J. and Thornton, J. M. (1996) Naccess version 2.1.1.

[51] Lee, B. and Richards, F. M. (1971) Interpretation of protein structures: estimation of static accessibility. *J.Mol.Biol. 55*, 379–400.

[52] Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. (2007) The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data. *Nucleic Acids Res 35*, D301–3.

[53] Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin,

J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., and Zardecki, C. (2002) The protein data bank. *Acta Crystallogr D Biol Crystallogr 58*, 899–907.

[54] Roche, O., Kiyama, R., and III, C. L. B. (2001) Ligand-protein database: Linking protein-ligand complex structures to binding data. *J.Med.Chem. 44*, 3592–3598.

[55] Chen, X., Lin, Y., and Gilson, M. K. (2001) The binding database: overview and user's guide. *Biopolymers 61*, 127–141.

[56] Wang, R., Fang, X., Lu, Y., and Wang, S. (2004) The pdbbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem 47*, 2977–80.

[57] Wang, R., Fang, X., Lu, Y., Yang, C.-Y., and Wang, S. (2005) The pdbbind database: methodologies and updates. *J Med Chem 48*, 4111–9.

[58] Golovin, A., Dimitropoulos, D., Oldfield, T., Rachedi, A., and Henrick, K. (2004) Msdsite: A database search and retrieval system for the analysis and viewing of bound ligands and active sites. *Proteins: Struct., Funct., Bioinf. 58*, 190–199.

[59] Bergner, A., Gunther, J., Hendlich, M., Klebe, G., and Verdonk, M. (2001) Use of r—elibase for retrieving complex three-dimensional interaction patterns including crystallographic packing effects. *Biopolymers 61*, 99–110.

[60] Hendlich, M., Bergner, A., Gunther, J., and Klebe, G. (2003) Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J Mol Biol 326*, 607–20.

[61] Puvanendrampillai, D. and Mitchell, J. B. O. (2003) L/d protein ligand database (pld): additional understanding of the nature and specificity of protein-ligand complexes. *Bioinformatics 19*, 1856–1857.

[62] Hu, L., Benson, M. L., Smith, R. D., Lerner, M. G., and Carlson, H. A. (2005) Binding moad (mother of all databases). *Proteins 60*, 333–340.

[63] Lane, D. P. (1992) Cancer. p53, guardian of the genome. *Nature 358*, 15–6.

[64] Levine, A. J. (1997) p53, the cellular gatekeeper for growth and division. *Cell 88*, 323–31.

[65] Vogelstein, B., Lane, D., and Levine, A. J. (2000) Surfing the p53 network. *Nature 408*, 307–10.

[66] Soussi, T., Dehouche, K., and Beroud, C. (2000) p53 website and analysis of p53 gene mutations in human cancer: forging a link between epidemiology and carcinogenesis. *Hum Mutat 15*, 105–13.

[67] Eymin, B., Gazzeri, S., Brambilla, C., and Brambilla, E. (2002) Mdm2 overexpression and p14(arf) inactivation are two mutually exclusive events in primary human lung tumors. *Oncogene 21*, 2750–61.

[68] Polsky, D., Bastian, B. C., Hazan, C., Melzer, K., Pack, J., Houghton, A., Busam, K., Cordon-Cardo, C., and Osman, I. (2001) Hdm2 protein overexpression, but not gene amplification, is related to tumorigenesis of cutaneous melanoma. *Cancer Res 61*, 7642–6.

[69] Momand, J., Jung, D., Wilczynski, S., and Niland, J. (1998) The mdm2 gene amplification database. *Nucleic Acids Res 26*, 3453–9.

[70] Chene, P. (2003) Inhibiting the p53-mdm2 interaction: an important target for cancer therapy. *Nat Rev Cancer 3*, 102–9.

[71] Vassilev, L. T. (2005) p53 activation by small molecules: application in oncology. *J Med Chem 48*, 4491–9.

[72] Kussie, P. H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A. J., and Pavletich, N. P. (1996) Structure of the mdm2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science 274*, 948–53.

[73] Barrett, C. P., Hall, B. A., and Noble, M. E. (2004) Dynamite: a simple way to gain insight into protein motions. *Acta Crystallogr D Biol Crystallogr 60*, 2280–7.

[74] Chen, H. F. and Luo, R. (2007) Binding induced folding in p53-mdm2 complex. *J Am Chem Soc 129*, 2930–7.

[75] Espinoza-Fonseca, L. M. and Trujillo-Ferrara, J. G. (2006) Conformational changes of the p53-binding cleft of mdm2 revealed by molecular dynamics simulations. *Biopolymers 83*, 365–73.

[76] Zhong, H. and Carlson, H. A. (2005) Computational studies and peptidomimetic design for the human p53-mdm2 complex. *Proteins 58*, 222–234.

[77] de Groot, B. L., van Aalten, D. M., Scheek, R. M., Amadei, A., Vriend, G., and Berendsen, H. J. (1997) Prediction of protein conformational freedom from distance constraints. *Proteins 29*, 240–51.

[78] Bowman, A. L., Nikolovska-Coleska, Z., Zhong, H., Wang, S., and Carlson, H. A. (2007) Small molecule inhibitors of the mdm2-p53 interaction discovered by ensemble-based receptor models. *J Am Chem Soc 129*, 12809–12814.

[79] Buolamwini, J. K., Addo, J., Kamath, S., Patil, S., Mason, D., and Ores, M. (2005) Small molecule antagonists of the mdm2 oncoprotein as anticancer agents. *Curr Cancer Drug Targets 5*, 57–68.

[80] Fischer, P. M. (2006) Peptide, peptidomimetic, and small-molecule antagonists of the p53-hdm2 protein-protein interaction. *Int J Pept Res Ther 12*, 3–19.

[81] Vassilev, L. T. (2007) Mdm2 inhibitors for cancer therapy. *Trends Mol Med 13*, 23–31.

[82] Vassilev, L. T., Vu, B. T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., Kong, N., Kammlott, U., Lukacs, C., Klein, C., Fotouhi, N., and Liu, E. A. (2004) In vivo activation of the p53 pathway by small-molecule antagonists of mdm2. *Science 303*, 844–8.

[83] Grasberger, B. L., Lu, T., Schubert, C., Parks, D. J., Carver, T. E., Koblish, H. K., Cummings, M. D., LaFrance, L. V., Milkiewicz, K. L., Calvo, R. R., Maguire, D., Lattanze, J., Franks, C. F., Zhao, S., Ramachandren, K., Bylebyl, G. R., Zhang, M., Manthey, C. L., Petrella, E. C., Pantoliano, M. W., Deckman, I. C., Spurlino, J. C., Maroney, A. C., Tomczuk, B. E., Molloy, C. J., and Bone, R. F. (2005) Discovery and cocrystal structure of benzodiazepinedione hdm2 antagonists that activate p53 in cells. *J Med Chem 48*, 909–12.

[84] McCoy, M. A., Gesell, J. J., Senior, M. M., and Wyss, D. F. (2003) Flexible lid to the p53-binding domain of human mdm2: implications for p53 regulation. *Proc Natl Acad Sci U S A 100*, 1645–8.

[85] Uhrinova, S., Uhrin, D., Powers, H., Watt, K., Zheleva, D., Fischer, P., McInnes, C., and Barlow, P. N. (2005) Structure of free mdm2 n-terminal domain reveals conformational adjustments that accompany p53-binding. *J Mol Biol 350*, 587–98.

[86] Shimizu, H. and Hupp, T. R. (2005) Intrasteric regulation of mdm2. *Trends Biochem Sci 28*, 346–349.

[87] Schmid, M. B. (2006) Crystallizing new approaches for antimicrobial drug discovery. *Biochem Pharmacol 71*, 1048–1056.

[88] Brewerton, S. C. (2008) The use of protein-ligand interaction fingerprints in docking. *Curr Opin Drug Discov Devel 11*, 356–364.

[89] Rester, U. (2006) Dock around the clock - current status of smoll molecule docking and scoring. *QSAR Comb Sci 25*, 606–615.

[90] Rognan, D. (2007) Chemogenomic approaches to rational drug design. *Br J Pharmacol 152*, 38–52.

[91] Kontoyianni, M., Madhav, P., Suchanek, E., and Seibel, W. (2008) Theoretical and practical considerations in virtual screening: a beaten field? *Curr Med Chem 15*, 107–116.

[92] Kirchmair, J., Markt, P., Distinto, S., Schuster, D., Spitzer, G. M., Liedl, K. R., Langer, T., and Wolber, G. (2008) The protein data bank (pdb), its related services and software tools as key components for in silico guided drug discovery. *J Med Chem 51*, 7021–7040.

[93] Block, P., Sotriffer, C. A., Dramburg, I., and Klebe, G. (2006) Affindb: a freely accessible database of affinities for protein-ligand complexes from the pdb. *Nucleic Acids Res 34*, D522–D526.

[94] Kellenberger, E., Muller, P., Schalon, C., Bret, G., Foata, N., and Rognan, D. (2006) sc-pdb: an annotated database of druggable binding sites from the protein data bank. *J Chem Inf Model 46*, 717–27.

[95] Bauer, R. A., Gnther, S., Jansen, D., Heeger, C., Thaben, P. F., and Preissner, R. (2009) Supersite: dictionary of metabolite and drug binding sites in proteins. *Nucleic Acids Res 37*, D195–D200.

[96] Li, L., Dantzer, J. J., Nowacki, J., O'Callaghan, B. J., and Meroueh, S. O. (2008) Pdbcal: a comprehensive dataset for receptor-ligand interactions with three-dimensional structures and binding thermodynamics from isothermal titration calorimetry. *Chem Biol Drug Des 71*, 529–532.

[97] Reddy, A. S., Amarnath, H. S. D., Bapi, R. S., Sastry, G. M., and Sastry, G. N. (2008) Protein ligand interaction database (plid). *Comput Biol Chem 32*, 387–390.

[98] Wallach, I. and Lilien, R. (2009) The protein-small-molecule database, a non-redundant structural resource for the analysis of protein-ligand binding. *Bioinformatics 25*, 615–620.

[99] Good, A. C. and Hermsmeier, M. A. (2007) Measuring camd technique performance. 2. how "druglike" are drugs? implications of random test set selection exemplified using druglikeness classification models. *J Chem Inf Model 47*, 110–114.

[100] Leach, A. R., Shoichet, B. K., and Peishoff, C. E. (2006) Prediction of protein-ligand interactions. docking and scoring: successes and gaps. *J Med Chem 49*, 5851–5.

[101] Minai, R., Matsuo, Y., Onuki, H., and Hirota, H. (2008) Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions. *Proteins 72*, 367–381.

[102] Irwin, J. J. (2008) Community benchmarks for virtual screening. *J Comput Aided Mol Des 22*, 193–199.

[103] Rosania, G. R., Crippen, G., Woolf, P., States, D., and Shedden, K. (2007) A cheminformatic toolkit for mining biomedical knowledge. *Pharm Res 24*, 1791–1802.

[104] Rhrig, U. F., Grosdidier, A., Zoete, V., and Michielin, O. (2009) Docking to heme proteins. *J Comput Chem 30*, 2305–2315.

[105] Park, K. and Kim, D. (2008) Binding similarity network of ligand. *Proteins 71*, 960–971.

[106] Daily, M. D. and Gray, J. J. (2007) Local motions in a benchmark of allosteric proteins. *Proteins 67*, 385–99.

[107] Kelley, L. A., Shrimpton, P. J., Muggleton, S. H., and Sternberg, M. J. E. (2009) Discovering rules for protein-ligand specificity using support vector inductive logic programming. *Protein Eng Des Sel 22*, 561–567.

[108] Saranya, N. and Selvaraj, S. (2009) Variation of protein binding cavity volume and ligand volume in protein-ligand complexes. *Bioorg Med Chem Lett 19*, 5769–5772.

[109] Zhou, P., Zou, J., Tian, F., and Shang, Z. (2009) Fluorine bonding–how does it work in protein-ligand interactions? *J Chem Inf Model 49*, 2344–2355.

[110] Odorico, M., Teulon, J.-M., Bessou, T., Vidaud, C., Bellanger, L., wen W Chen, S., Qumneur, E., Parot, P., and Pellequer, J.-L. (2007) Energy landscape of chelated uranyl: antibody interactions by dynamic force spectroscopy. *Biophys J 93*, 645–654.

[111] Wells, J. A. and McClendon, C. L. (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature 450*, 1001–1009.

[112] Xin, Y., Gadda, G., and Hamelberg, D. (2009) The cluster of hydrophobic residues controls the entrance to the active site of choline oxidase. *Biochemistry 48*, 9599–9605.

[113] Warr, W. (2009) Fragment-based drug discovery. *J Comput Aided Mol Des.*

[114] Böhm, H. J. (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des 8*, 243–256.

[115] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The protein data bank. *Nucleic Acids Res. 28*, 235–242.

[116] (year?) Information about pdbbeta's ligand explorer can be found at http://pdbbeta.rcsb.org/pdb/welcome.do.

[117] Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H. M., and Westbrook, J. (2004) Ligand depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics 20*, 2153–5.

[118] Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L., and Thornton, J. M. (1997) Pdbsum: a web-based database of summaries and analyses of all pdb structures. *Trends Biochem.Sci. 22*, 488–490.

[119] Luscombe, N. M., Laskowski, R. A., Westhead, D. R., Milburn, D., Jones, S., Karmirantzou, M., and Thornton, J. M. (1998) New tools and resources for analysing protein structures and their interactions. *Acta Crystallogr D Biol Crystallogr 54*, 1132–1138.

[120] Laskowski, R. A. (2001) Pdbsum: summaries and analyses of pdb structures. *Nucleic Acids Res. 29*, 221–222.

[121] Chen, J., Anderson, J. B., DeWeese-Scott, C., Fedorova, N. D., Geer, L. Y., He, S., Hurwitz, D. I., Jackson, J. D., Jacobs, A. R., Lanczycki, C. J., Liebert, C. A., Liu, C., Madej, T., Marchler-Bauer, A., Marchler, G. H., Mazumder, R., Nikolskaya, A. N., Rao, B. S., Panchenko, A. R., Shoemaker, B. A., Simonyan,

V., Song, J. S., Thiessen, P. A., Vasudevan, S., Wang, Y., Yamashita, R. A., Yin, J. J., and Bryant, S. H. (2003) Mmdb: Entrez's 3d-structure database. *Nucleic Acids Res. 31*, 474–477.

[122] Rockey, W. M. and Elcock, A. H. (2002) Progress toward virtual screening for drug side effects. *Proteins 48*, 664–671.

[123] Paul, N., Kellenberger, E., Bret, G., Mueller, P., and Rognan, D. (2004) Recovering the true targets of specific ligands by virtual screening of the protein data bank. *Proteins: Struct., Funct., Bioinf. 54*, 671–680.

[124] Westbrook, J. D. and Bourne, P. E. (2000) Star/mmcif: an ontology for macromolecular structure. *Bioinformatics 16*, 159–168.

[125] Westbrook, J., Feng, Z., Jain, S., Bhat, T. N., Thanki, N., Ravichandran, V., Gilliland, G. L., Bluhm, W. F., Weissig, H., Greer, D. S., Bourne, P. E., and Berman, H. M. (2002) The protein data bank: unifying the archive. *Nucleic Acids Res. 30*, 245–248.

[126] Christianson, D. W. (1991) Structural biology of zinc. *Adv.Protein Chem. 42*, 281–355.

[127] Verras, A., Kuntz, I. D., and Ortiz de Montellano, P. R. (2004) Computer-assisted design of selective imidazole inhibitors for cytochrome p450 enzymes. *J.Med.Chem. 47*, 3572–3579.

[128] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res. 25*, 3389–3402.

[129] Stebbings, L. A. and Mizuguchi, K. (2004) Homstrad: recent developments of

the homologous protein structure alignment database. *Nucleic Acids Res. 32*, D203–D207.

[130] Weissig, H. and Bourne, P. E. (2002) Protein structure resources. *Acta Crystallogr D Biol Crystallogr 58*, 908–15.

[131] Lamb, M. L. (2005) Targeting the kinome with computational chemistry. *Annu.Rep.Comput.Chem. 1*, 185–202.

[132] Mao, L., Wang, Y., Liu, Y., and Hu, X. (2004) Molecular determinants for atp-binding in proteins: A data mining and quantum chemical analysis. *J.Mol.Biol. 336*, 787–807.

[133] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2001) *GATE: an architecture for development of robust HLT applications.* (Association for Computational Linguistics, Morristown, NJ, USA), pp. 168–175. inproceedings mbensonz.

[134] Deshpande, N., Addess, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z., Green, R. K., Flippen-Anderson, J. L., Westbrook, J., Berman, H. M., and Bourne, P. E. (2005) The rcsb protein data bank: a redesigned query system and relational database based on the mmcif schema. *Nucleic Acids Res. 33*, D233–D237.

[135] Laskowski, R. A., Chistyakov, V. V., and Thornton, J. M. (2005) Pdbsum more: New summaries and analyses of the known 3d structures of proteins and nucleic acids. *Nucleic Acids Res. 33*, D266–D268.

[136] Binkowski, T. A., Adamian, L., and Liang, J. (2003) Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J.Mol.Biol. 332*, 505–526.

[137] Shin, J.-M. and Cho, D.-H. (2005) Pdb-ligand: a ligand database based on pdb for the automated and customized classification of ligand-binding structures. *Nucleic Acids Res 33*, D238–41.

[138] Hendlich, M. (1998) Databases for protein-ligand complexes. *Acta Crystallogr D Biol Crystallogr 54*, 1178–1182.

[139] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003) The swiss-prot protein knowledge-base and its supplement trembl in 2003. *Nucleic Acids Res. 31*, 365–370.

[140] Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A., and Rapp, B. A. (2000) Database resources of the national center for biotechnology information. *Nucleic Acids Res. 28*, 10–14.

[141] Delano, W. L. (2002) The pymol molecular graphics system, delano scientific llc (San Carlos, CA, USA, http://www.pymol.org).

[142] Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V., and Subramaniam, S. (1998) Analytical shape computation of macromolecules: Ii. inaccessible cavities in proteins. *Proteins: Struct., Funct., Genet. 33*, 18–29.

[143] Richards, F. M. (1985) Calculation of molecular volumes and areas for structures of known geometry. *Methods Enzymol. 115*, 440–464.

[144] Ho, C. M. and Marshall, G. R. (1990) Cavity search: an algorithm for the isolation and display of cavity-like binding regions. *J Comput Aided Mol Des 4*, 337–354.

[145] Goede, A., Preissner, R., and Froemmel, C. (1997) Voronoi cell: new method for

allocation of space among atoms: elimination of avoidable errors in calculation of atomic volume and density. *J.Comput.Chem. 18*, 1113–1123.

[146] McConkey, B. J., Sobolev, V., and Edelman, M. (2002) Quantification of protein surfaces, volumes and atom-atom contacts using a constrained voronoi procedure. *Bioinformatics 18*, 1365–1373.

[147] Fleming, P. J. and Richards, F. M. (2000) Protein packing: Dependence on protein size, secondary structure and amino acid composition. *J.Mol.Biol. 299*, 487–498.

[148] Poupon, A. (2004) Voronoi and voronoi-related tessellations in studies of protein structure and interaction. *Curr.Opin.Struct.Biol. 14*, 233–241.

[149] Stahl, M., Taroni, C., and Schneider, G. (2000) Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng. 13*, 83–88.

[150] Jorgensen, W. L. and Tirado-Rives, J. (1988) The opls [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J.Am.Chem.Soc. 110*, 1657–1666.

[151] Li, A.-J. and Nussinov, R. (1998) A set of van der waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *Proteins: Struct., Funct., Genet. 32*, 111–127.

[152] Tsai, J., Taylor, R., Chothia, C., and Gerstein, M. (1999) The packing density in proteins: Standard radii and volumes. *J.Mol.Biol. 290*, 253–266.

[153] Chothia, C. (1976) The nature of the accessible and buried surfaces in proteins. *J.Mol.Biol. 105*, 1–12.

[154] An, J., Totrov, M., and Abagyan, R. (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol.Cell.Proteomics 4*, 752–761.

[155] DesJarlais, R. L., Sheridan, R. P., Seibel, G. L., Dixon, J. S., Kuntz, I. D., and Venkataraghavan, R. (1988) Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J.Med.Chem. 31*, 722–729.

[156] Laskowski, R. A., Luscombe, N. M., Swindells, M. B., and Thornton, J. M. (1996) Protein clefts in molecular recognition and function. *Protein Sci. 5*, 2438–2452.

[157] Ferre, F., Ausiello, G., Zanzoni, A., and Helmer-Citterich, M. (2004) Surface: a database of protein surface regions for functional annotation. *Nucleic Acids Res 32*, D240–4.

[158] Gutteridge, A., Bartlett, G. J., and Thornton, J. M. (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J Mol Biol 330*, 719–34.

[159] Hofbauer, C., Lohninger, H., and Aszodi, A. (2004) Surfcomp: a novel graph-based approach to molecular surface comparison. *J Chem Inf Comput Sci 44*, 837–47.

[160] Ivanisenko, V. A., Pintus, S. S., Grigorovich, D. A., and Kolchanov, N. A. (2005) Pdbsite: A database of the 3d structure of protein functional sites. *Nucleic Acids Res. 33*, D183–D187.

[161] Keil, M., Exner, T. E., and Brickmann, J. (2004) Pattern recognition strategies for molecular surfaces: Iii. binding site prediction with a neural network. *J.Comput.Chem. 25*, 779–789.

[162] Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. (2003) Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A 100*, 5772–7.

[163] Betts, M. J. and Sternberg, M. J. E. (1999) An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng. 12*, 271–283.

[164] Fradera, X., de la Cruz, X., Silva, C. H. T. P., Gelpi, J. L., Luque, F. J., and Orozco, M. (2002) Ligand-induced changes in the binding sites of proteins. *Bioinformatics 18*, 939–948.

[165] Halle, B. (2002) Flexibility and packing in proteins. *Proc Natl Acad Sci U S A 99*, 1274–9.

[166] Najmanovich, R., Kuttner, J., Sobolev, V., and Edelman, M. (2000) Side-chain flexibility in proteins upon ligand binding. *Proteins: Struct., Funct., Genet. 39*, 261–268.

[167] Zhao, S., Goodsell, D. S., and Olson, A. J. (2001) Analysis of a data set of paired uncomplexed protein structures: new metrics for side-chain flexibility and model evaluation. *Proteins: Struct., Funct., Genet. 43*, 271–279.

[168] Bhm, H. J. and Stahl, M. (2002) *The use of scoring functions in drug discovery applications, in: K.B. Lipkowitz, D.B. Boyd (Eds). Reviews in Computational Chemistry vol. 18.* (Wiley-VCH Inc.), pp. 41–88.

[169] Smith, R. D., Hu, L., Falkner, J. A., Benson, M. L., Nerothin, J. P., and Carlson, H. A. (2006) Exploring protein-ligand recognition with binding moad. *J Mol Graph Model 24*, 414–425.

[170] Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev 46*, 3–26.

[171] Sugiyama, Y. (2005) Druggability: selecting optimized drug candidates. *Drug Discov Today 10*, 1577–9.

[172] Norvell, J. C. and Machalek, A. Z. (2000) Structural genomics programs at the us national institute of general medical sciences. *Nat Struct Biol 7 Suppl*, 931.

[173] Luque, I. and Freire, E. (1998) Structure-based prediction of binding affinities and molecular design of peptide ligands. *Methods Enzymol 295*, 100–127.

[174] Williams, D. H., Stephens, E., O'Brien, D. P., and Zhou, M. (2004) Understanding noncovalent interactions: ligand binding energy and catalytic efficiency from ligand-induced reductions in motion within receptors and enzymes. *Angew Chem Int Ed Engl 43*, 6596–6616.

[175] Coleman, R. G., Salzberg, A. C., and Cheng, A. C. (2006) Structure-based identification of small molecule binding sites using a free energy model. *J Chem Inf Model 46*, 2631–7.

[176] Hajduk, P. J., Huth, J. R., and Tse, C. (2005) Predicting protein druggability. *Drug Discov Today 10*, 1675–82.

[177] (2007) Molecular operating environment (moe), 2007.08.

[178] Wildman, S. A. and Crippen, G. M. (1999) Prediction of physicochemical parameters by atomic contributions. *J Chem Inf Comput Sci 39*, 868–873.

[179] (2002) Sas, release 9.1.

[180] (2007) Jmp, release 7.01.

[181] Coleman, R. G. and Sharp, K. A. (2006) Travel depth, a new shape descriptor for macromolecules: application to ligand binding. *J Mol Biol 362*, 441–458.

[182] Yang, C.-Y., Wang, R., and Wang, S. (2006) M-score: a knowledge-based potential scoring function accounting for protein atom mobility. *J Med Chem 49*, 5903–5911.

[183] Babaoglu, K. and Shoichet, B. K. (2006) Deconstructing fragment-based inhibitor discovery. *Nat Chem Biol 2*, 720–3.

[184] Carr, R., Congreve, M., Murray, C., and Rees, D. (2005) Fragment-based lead discovery: leads by design. *Drug Discov Today 10*, 987–992.

[185] Hajduk, P. J. (2006) Fragment-based drug design: how big is too big? *J Med Chem 49*, 6972–6976.

[186] Chothia, C. (1974) Hydrophobic bonding and accessible surface area in proteins. *Nature 248*, 338–339.

[187] DeYoung, L. and Dill, K. (1990) Partitioning of nonpolar solutes into bilayers and amorphous n-alkanes. *J Phys Chem 94*, 801–809.

[188] An, J., Totrov, M., and Abagyan, R. (2004) Comprehensive identification of "druggable" protein ligand binding sites. *Genome Inform 15*, 31–41.

[189] Hopkins, A. L. and Groom, C. R. (2002) The druggable genome. *Nat Rev Drug Discov 1*, 727–730.

[190] Kubinyi, H. (2003) Drug research: myths, hype and reality. *Nat Rev Drug Discov 2*, 665–668.

[191] Strachan, R. T., Ferrara, G., and Roth, B. L. (2006) Screening the receptorome: an efficient approach for drug discovery and target validation. *Drug Discov Today 11*, 708–716.

[192] Whitty, A. and Kumaravel, G. (2006) Between a rock and a hard place? *Nat Chem Biol 2*, 112–118.

[193] Thanos, C. D., DeLano, W. L., and Wells, J. A. (2006) Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc Natl Acad Sci U S A 103*, 15422–15427.

[194] Russ, A. P. and Lampel, S. (2005) The druggable genome: an update. *Drug Discov Today 10*, 1607–1610.

[195] Hajduk, P. J., Huth, J. R., and Fesik, S. W. (2005) Druggability indices for protein targets derived from nmr-based screening data. *J Med Chem 48*, 2518–25.

[196] Bogan, A. A. and Thorn, K. S. (1998) Anatomy of hot spots in protein interfaces. *J Mol Biol 280*, 1–9.

[197] DeLano, W. L. (2002) Unraveling hot spots in binding interfaces: progress and challenges. *Curr Opin Struct Biol 12*, 14–20.

[198] Soga, S., Shirai, H., Kobori, M., and Hirayama, N. (2007) Use of amino acid composition to predict ligand-binding sites. *J Chem Inf Model 47*, 400–6.

[199] Cheng, A. C., Coleman, R. G., Smyth, K. T., Cao, Q., Soulard, P., Caffrey, D. R., Salzberg, A. C., and Huang, E. S. (2007) Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol 25*, 71–75.

[200] Brooijmans, N. and Kuntz, I. D. (2003) Molecular recognition and docking algorithms. *Annu Rev Biophys Biomol Struct 32*, 335–373.

[201] Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins 47*, 409–443.

[202] Krovat, E., Steindle, T., and Langer, T. (2005) Recent advances in docking and scoring. *Curr ComputAided-Drug-Des 1*, 93–102.

[203] Mohan, V., Gibbs, A. C., Cummings, M. D., Jaeger, E. P., and DesJarlais, R. L. (2005) Docking: successes and challenges. *Curr Pharm Des 11*, 323–33.

[204] Mobley, D. L., Graves, A. P., Chodera, J. D., McReynolds, A. C., Shoichet, B. K., and Dill, K. A. (2007) Predicting absolute ligand binding free energies to a simple model site. *J Mol Biol 371*, 1118–1134.

[205] Marsden, P. M., Puvanendrampillai, D., Mitchell, J. B. O., and Glen, R. C. (2004) Predicting protein-ligand binding affinities: a low scoring game? *Org Biomol Chem 2*, 3267–3273.

[206] Nissink, J. W. M. (2009) Simple size-independent measure of ligand efficiency. *J Chem Inf Model 49*, 1617–1622.

[207] Reynolds, C. H., Bembenek, S. D., and Tounge, B. A. (2007) The role of molecular size in ligand efficiency. *Bioorg Med Chem Lett 17*, 4258–4261.

[208] Reynolds, C. H., Tounge, B. A., and Bembenek, S. D. (2008) Ligand binding efficiency: trends, physical basis, and implications. *J Med Chem 51*, 2432–2438.

[209] (2007) Scitegic pipeline pilot, 6.1.5.0 (Accelrys, San Diego, CA 92121).

[210] Carlson, H. A., Smith, R. D., Khazanov, N. A., Kirchhoff, P. D., Dunbar, J. B., and Benson, M. L. (2008) Differences between high- and low-affinity complexes of enzymes and nonenzymes. *J Med Chem 51*, 6432–6441.

[211] Verdonk, M. L. and Rees, D. C. (2008) Group efficiency: a guideline for hits-to-leads chemistry. *ChemMedChem 3*, 1179–1180.

[212] Keser, G. M. and Makara, G. M. (2009) The influence of lead discovery strategies on the properties of drug candidates. *Nat Rev Drug Discov 8*, 203–212.

[213] Zhang, X. and Houk, K. N. (2005) Why enzymes are proficient catalysts: beyond the pauling paradigm. *Acc Chem Res 38*, 379–385.

[214] Houk, K. N., Leach, A. G., Kim, S. P., and Zhang, X. (2003) Binding affinities of host-guest, protein-ligand, and protein-transition-state complexes. *Angew Chem Int Ed Engl 42*, 4872–4897.

[215] Brooijmans, N., Sharp, K. A., and Kuntz, I. D. (2002) Stability of macromolecular complexes. *Proteins 48*, 645–53.

[216] Deremble, C. and Lavery, R. (2005) Macromolecular recognition. *Curr Opin Struct Biol 15*, 171–175.

[217] Zhang, E., Hatada, M., Brewer, J. M., and Lebioda, L. (1994) Catalytic metal ion binding in enolase: the crystal structure of an enolase-mn2+-phosphonoacetohydroxamate complex at 2.4-a resolution. *Biochemistry 33*, 6295–6300.

[218] Kurinov, I. V. and Harrison, R. W. (1994) Prediction of new serine proteinase inhibitors. *Nat Struct Biol 1*, 735–743.

[219] Paesen, G. C., Adams, P. L., Harlos, K., Nuttall, P. A., and Stuart, D. I. (1999) Tick histamine-binding proteins: isolation, cloning, and three-dimensional structure. *Mol Cell 3*, 661–671.

[220] Li, H., Raman, C. S., Martsek, P., Krl, V., Masters, B. S., and Poulos, T. L. (2000) Mapping the active site polarity in structures of endothelial nitric oxide synthase heme domain complexed with isothioureas. *J Inorg Biochem 81*, 133–139.

[221] Fischmann, T. O., Hruza, A., Niu, X. D., Fossetta, J. D., Lunn, C. A., Dolphin, E., Prongay, A. J., Reichert, P., Lundell, D. J., Narula, S. K., and Weber,

P. C. (1999) Structural characterization of nitric oxide synthase isoforms reveals striking active-site conservation. *Nat Struct Biol 6*, 233–242.

[222] Mller, A., Thomas, G. H., Horler, R., Brannigan, J. A., Blagova, E., Levdikov, V. M., Fogg, M. J., Wilson, K. S., and Wilkinson, A. J. (2005) An atp-binding cassette-type cysteine transporter in campylobacter jejuni inferred from the structure of an extracytoplasmic solute receptor protein. *Mol Microbiol 57*, 143–155.

[223] Vassylyev, D. G., Tomitori, H., Kashiwagi, K., Morikawa, K., and Igarashi, K. (1998) Crystal structure and mutational analysis of the escherichia coli putrescine receptor. structural basis for substrate specificity. *J Biol Chem 273*, 17604–17609.

[224] Benini, S., Rypniewski, W. R., Wilson, K. S., Miletti, S., Ciurli, S., and Mangani, S. (2000) The complex of bacillus pasteurii urease with acetohydroxamate anion from x-ray data at 1.55 a resolution. *J Biol Inorg Chem 5*, 110–118.

[225] Cameron, A., Read, J., Tranter, R., Winter, V. J., Sessions, R. B., Brady, R. L., Vivas, L., Easton, A., Kendrick, H., Croft, S. L., Barros, D., Lavandera, J. L., Martin, J. J., Risco, F., Garca-Ochoa, S., Gamo, F. J., Sanz, L., Leon, L., Ruiz, J. R., Gabarr, R., Mallo, A., and de las Heras, F. G. (2004) Identification and activity of a series of azole-based compounds with lactate dehydrogenase-directed anti-malarial activity. *J Biol Chem 279*, 31429–31439.

[226] Oh, B. H., Ames, G. F., and Kim, S. H. (1994) Structural basis for multiple ligand specificity of the periplasmic lysine-, arginine-, ornithine-binding protein. *J Biol Chem 269*, 26323–26330.

[227] Inanobe, A., Furukawa, H., and Gouaux, E. (2005) Mechanism of partial agonist action at the nr1 subunit of nmda receptors. *Neuron 47*, 71–84.

[228] Furukawa, H. and Gouaux, E. (2003) Mechanisms of activation, inhibition and specificity: crystal structures of the nmda receptor nr1 ligand-binding core. *EMBO J 22*, 2873–2885.

[229] Bjrkman, A. J., Binnie, R. A., Zhang, H., Cole, L. B., Hermodson, M. A., and Mowbray, S. L. (1994) Probing protein-protein interactions. the ribose-binding protein in bacterial transport and chemotaxis. *J Biol Chem 269*, 30206–30211.

[230] Kangas, E. and Tidor, B. (1998) Optimizing electrostatic affinity in ligand-receptor: Theory computation and ligand properties. *J Chem Phys 109*, 7522–7545.

[231] Kumar, S. and Nussinov, R. (2002) Close-range electrostatic interactions in proteins. *Chembiochem 3*, 604–17.

[232] Hendsch, Z. S. and Tidor, B. (1994) Do salt bridges stabilize proteins? a continuum electrostatic analysis. *Protein Sci 3*, 211–26.

[233] Musafia, B., Buchner, V., and Arad, D. (1995) Complex salt bridges in proteins: statistical analysis of structure and function. *J Mol Biol 254*, 761–770.

[234] Olson, C. A., Spek, E. J., Shi, Z., Vologodskii, A., and Kallenbach, N. R. (2001) Cooperative helix stabilization by complex arg-glu salt bridges. *Proteins 44*, 123–132.

[235] Kumar, S. and Nussinov, R. (1999) Salt bridge stability in monomeric proteins. *J Mol Biol 293*, 1241–1255.

[236] Tan, Z. J. and Chen, S. J. (2005) Electrostatic correlations and fluctuations for ion binding to a finite length polyelectrolyte. *J Chem Phys 122*, 44903.

[237] Stanley, C. and Rau, D. C. (2006) Preferential hydration of dna: the magnitude

and distance dependence of alcohol and polyol interactions. *Biophys J 91*, 912–920.

[238] Schitt, B., Iversen, B. B., Madsen, G. K., Larsen, F. K., and Bruice, T. C. (1998) On the electronic nature of low-barrier hydrogen bonds in enzymatic reactions. *Proc Natl Acad Sci U S A 95*, 12799–12802.

[239] Cleland, W. W. and Kreevoy, M. M. (1994) Low-barrier hydrogen bonds and enzymic catalysis. *Science 264*, 1887–1890.

[240] Haaland, A. (1989) Covalent versus dative bonds to main group metals. *Angew Chem Int Ed Engl 1989*, 992–1007.

[241] Corzo, J. (2006) Time the forgotten dimension of ligand binding teaching. *Biochem Mol Biol Edu 34*, 413–416.

[242] Espinoza-Fonseca, L. M. and Garca-Machorro, J. (2008) Aromatic-aromatic interactions in the formation of the mdm2-p53 complex. *Biochem Biophys Res Commun 370*, 547–551.

[243] Zhao, J., Wang, M., Chen, J., Luo, A., Wang, X., Wu, M., Yin, D., and Liu, Z. (2002) The initial evaluation of non-peptidic small-molecule hdm2 inhibitors based on p53-hdm2 complex structure. *Cancer Lett 183*, 69–77.

[244] Ding, K., Lu, Y., Nikolovska-Coleska, Z., Wang, G., Qiu, S., Shangary, S., Gao, W., Qin, D., Stuckey, J., Krajewski, K., Roller, P. P., and Wang, S. (2006) Structure-based design of spiro-oxindoles as potent, specific small-molecule inhibitors of the mdm2-p53 interaction. *J Med Chem 49*, 3432–3435.

[245] Showalter, S. A., Bruschweiler-Li, L., Johnson, E., Zhang, F., and Brschweiler, R. (2008) Quantitative lid dynamics of mdm2 reveals differential ligand binding modes of the p53-binding cleft. *J Am Chem Soc 130*, 6472–6478.

[246] Pearlman, D A amd Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham III, T. E., DeBolt, S., Ferguson, D. M., Seibel, G. L., and Kollman, P. A. (1995) Amber, a package of computer programs for applying molecular mechanics, normal-mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput Phys Comm 91*, 1–41.

[247] Case, D. A., Darden, T. A., Cheatham III, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Merz, K. M., Wang, B., Pearlman, D. A., Crowley, M., Brozell, S., Tsui, V., Gohlke, H., Mongan, J., Hornak, V., Cui, G., Beroza, P., Schafmeister, C., Caldwell, J W Ross, W. S., and Kollman, P. A. (2004) Amber8; university of california, san francisco: San francisco, ca.

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

[249] Wang, J., Wang, W., Kollman, P. A., and Case, D. A. (2006) Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model 25*, 247–260.

[250] Jakalian, A., Bush, B., Jack, D. B., and Bayly, C. I. (2000) Fast, efficient generation of high-quality atomic charges. am1-bcc model: I. method. *J Comput Chem 21*, 132–136.

[251] Onufriev, A., Bashford, D., and Case, D. A. (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins 55*, 383–394.

[252] Damm, K. L. and Carlson, H. A. (2006) Gaussian-weighted rmsd superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. *Biophys J 90*, 4558–4573.

[253] Feig, M., Karanicolas, J., and Brooks, C. L. (2001) Mmtsb tool set; the scripps research institute: La jolla, ca.