Binding MOAD
(Mother of All Databases)

by

Mark Benson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2009

Doctoral Committee:

Associate Professor Heather A. Carlson, Chair
Professor Brian D. Athey
Professor Gordon M. Crippen
Professor Hosagrahar V. Jagadish
Professor Gilbert S. Omenn
Associate Professor Mark A. Saper

To my eternal companion, Rebekah, and my wonderful kids Jacob, Hannah and Emily

Acknowledgments

I would like to acknowledge the boundless encouragement and support of Dr. Heather Carlson. I express humble gratitude for being able to work in her lab and learn about many of the aspects of being a scientist. Her dedication and passion is examplar. I am thankful for the many opportunities that she has provided for continued growth, and her sacrifices made to help others. Working in her lab has been wonderful, I am proud to be a HACer. I also humbly thank my dissertation committee for their selfless contributions and recommendations.

I would like to thank the Carlson lab, for help, discussions, insight and patience. Without Richard Smith, Binding MOAD would not have been possible. Nickolay Khazanov and Jim Dunbar and Michael Lerner have been great help in getting work done and having critical discussions. I have learned a lot about programming from both Jayson Falkner and Jason Nerothin. I acknowledge Derek Mendez, Liegi Hu, Steve Spronk, Katrina Lexa, Man-Un (Peter) Ung, Anna Bowman, Jerome Quintero, Kristin Meagher, Kelly Damm, Joslyn Kravitz, Xiao-Jian Tan, and Paul D. Kirchhoff. I humbly thank Lynn Alexander, Julia Eussen, and Yuri Santos for all their endless service. Allan Bailey has been a saint, kept the machines steadily humming and lights blinking.

I would also like to thank those at Torrey Path, LLC (Metamatics, LLC) for their help and contributions. Their help with BUDA has save countless hours spent in the annual process of adding data to Binding MOAD. Namely, thanks goes to John Beaver, Brandon Dimcheff, and Peter Dresslar.

Open source tools from several organizations have been of great value in this work. Namely, I wish to acknowledge MySQL AB for their MySQL database, RedHat for the JBoss application Server, the Apache Project for Jakarta Struts, Novel for SuSE Linux, Warren DeLano for PyMOL, and ChemAxon for their cheminformatic web-based tools.

I thank the Chemical Computing Group, Inc. for their generous donation of MOE for calculating the physical properties of the ligands. I also thank the Center for Statistical Consultation and Research for invaluable help with statistical analysis.

Chapter 2 has been adapted from two previously published papers

Benson, M.L., Smith, R.D., Khazanov, N.A., Dimcheff. B., Beaver, J., Dresslar. P., Nerothin. J., and Carlson, H.A. Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res* **2008**, *36(Database issue)*, D674-8.

and

Hu, L., Benson, M.L., Smith, R.D., Lerner, M.G., and Carlson, H.A. Binding MOAD (Mother Of All Databases). *Proteins*. **2005**, *60*, 333-40.

Chapter 3 has been previously published as

Carlson, H.A., Smith, R.D., Khazanov, N.A., Kirchhoff, P.D., Dunbar, J.B. Jr, and Benson, M.L. Differences between high- and low-affinity complexes of enzymes and nonenzymes. *J Med Chem* **2008**, *51*, 6432-41.

Appendix B has been accepted to be published as

Benson, M.L, Smith, R.D., Khazanov, N.A., Dimcheff, B.C., Dresslar, P., and Carlson, H.A. Updating Binding MOAD - Data Management and Information Workflow *New Mathematics and Natural Computation*

This work was funded by an NSF CAREER Award (MCB-0546073), the National Institutes of Health (HG003890) and a Beckman Young Investigator award to HAC.

Table of Contents

List of Tables

List of Figures

Figure

List of Appendices

Appendix

Abstract

Binding MOAD (Mother of All Databases) is the largest collection of high-quality, protein-ligand complexes available from the Protein Data Bank. At this time, Binding MOAD contains 11,368 protein-ligand complexes composed of 3583 unique protein families and 5363 unique ligands. We have searched the crystallography papers for all structures and compiled binding data for 3543 (31%) of the protein-ligand complexes. The binding-affinity data ranges 13 orders of magnitude. This is the largest collection of structural binding data to date in the literature.

This database of protein-ligand complexes is proving very useful in exploring biophysical patterns of molecular recognition and enzymatic regulation. Mining Binding MOAD has revealed physical differences in how enzymes and nonenzymes bind small molecules. High-affinity ligands of enzymes are much larger than those with low affinity, but high- and low-affinity ligands of nonenzymes are the same size. This suggests that different approaches may be appropriate for improving the affinity of ligands. While the addition of complementary functional groups is likely to improve the affinity of an enzyme inhibitor, it may not be as fruitful for ligands of nonenzymes. For nonenzymes, small changes and isosteric replacements might be more productive. Furthermore, nonenzymes were found to have higher ligand efficiencies. The different efficiencies are not due to differences in the physicochemical properties of the ligands; instead, the amino-acid composition of the pockets are very different despite very similar distributions of amino acids in the overall protein sequences.

This study aims to address the issue of protein flexibility upon ligand binding. The influence of ligand binding on protein flexibility is examined by analyzing a large number of proteins crystallized with and without ligands. A baseline comparison of the natural variation of protein structure with and without ligands is first established, and then differences between the apo and holo are analyzed. It is shown that, in general, ligand binding stabilizes the protein and results in a smaller backbone root mean square deviation (RMSD) among holo-protein structures, compared the backbone RMSD of the apo-protein structures. Furthermore, the holo structures appear to sample a smaller subset of the space inhabited

by apo structures, because the difference between apo and holo structures is smaller than variation seen among apo structures themselves. The size of the bound ligand does not appear to matter in determining the rigidification. While ligand binding generally does not induce large changes in the backbone, they are significant. Ligand binding does have distinct impact on the active site, as revealed by all-atom, active-site RMSD and the range of χ_1 variation. Apo structures are observed to have a certain range of flexibility in their active sites, just as holo structures have a similar, but smaller, degree of variation among their active sites. However, greater variation has been found between these two groups as opposed to within either group by themselves. This suggests that ligand binding induces active-site side chains to occupy a different conformational space before and after binding. The influence on the active site could not be easily attributed to features such as ligand size, resolution, protein function, or catalytic composition.

The studies above illustrate the usefulness of large carefully annotated datasets for studying protein-ligand interactions. Binding MOAD has almost doubled in size since it was originally introduced in 2004, demonstrating steady growth with each annual update. Several technologies are described, such as natural language processing, that help drive this constant expansion.

In summary, Binding MOAD is a valuable resource. It has helped to illuminate fundamental differences between enzymes and nonenzymes and allowed for examination of the influence ligand binding has in protein flexibility. It has great potential to further advance our understanding of protein-ligand interactions.

Chapter 1

Introduction

Proteins are responsible for many functions required for cellular life, such as signal transduction, metabolism, and structure. Many of these functions involve binding of small molecules. While these are small organic compounds, peptides, or nucleotides, they can, as a whole, be referred to as ligands, whether or not they are catalytic substrates.

Studying protein-ligand binding has important application in structure-based drug design (SBDD). A strategy in SBDD is to use computational techniques to virtually screen a database of potential drugs and identify which compounds are likely to be active and which compounds are likely to be inactive, thereby focusing and speeding up the drug development process. While virtual screening is proving very helpful to pharmaceutical research, being able to consistently and accurately predict activity of a compound is proving very difficult. [3, 4] Furthermore, protein-ligand binding is proving to be an essential component of fields such as chemical biology and metabolomics.

This dissertation studies protein-ligand binding by using an extensive database of high quality protein-ligand crystal structures called Binding MOAD (Mother Of All Databases). It uses this database to discover new principles of protein-ligand binding and also to refine existing hypothesis involved in ligand docking and scoring. This dissertation describes three primary studies. First, details of the database creation are described, including the criteria used in selecting protein-ligand complexes for inclusion and what properties the entries have as a collection. Second, a mining study searches for causes of highly efficient ligand binding. Third, protein flexibility upon ligand binding is probed. Lastly, this dissertation describes data structure and software development necessary for distributing Binding MOAD as well as software developed to facilitate the annual updating and annotation of data for Binding MOAD. As a whole, this dissertation outlines Binding MOAD as a major bioinformatics resource and uses it to provide useful insights into protein-ligand recognition.

1.1 Protein-Ligand Binding Theory

Theories about protein-ligand binding are constantly evolving. In 1894, Hermann Emil Fischer proposed the "lock and key" model to describe enzyme substrate interactions, "enzyme and glycoside must fit together like a key and a lock in order to initiate a chemical action upon each other."[5, 6, 7] Linus Pauling said in 1948, "I think that enzymes are molecules that are complementary in structure to the activated complexes of the reactions that they catalyze."[8, 9] Later, in 1958, Daniel Koshland established the "induced-fit" theory that the protein's binding site would adjust to accommodate the ligand.[10] The current theory described proteins existing in an equilibrium ensemble of energetically similar conformations.[11, 12, 13, 14, 15] A ligand may bind to any of the protein conformations. This binding may shift the equilibrium of the system to a new distribution favorable to the binding reaction.[16] Induced fit may be described as moderate binding to the lowest energy conformation or tight binding to a higher energy conformation, whereas lock-and-key binding could be described as tight binding to the lowest energy conformation.[11]

Ligand binding may induce protein structural rearrangements. These structural rearrangements may range from small side-chain movements to large loop reorganizations and domain hinge movements.

1.2 Protein Flexibility

Proteins are naturally flexible macromolecules. This flexibility comes from being composed of a string of amino acids, folded into a structure stabilized by non-covalent interactions. Flexibility is an important component of binding substrates, catalyzing enzymatic reactions, and releasing the products.[15, 17]

Studying the flexibility of binding sites has yielded some key insights. Jaap Heringa and Patrick Argos observed cases of ligand binding induced strain in clusters of residues in ligand binding sites. The strain was manifested as nonrotameric side-chain positions with tight packing.[18] They hypothesized that the strain of displacing the side chain from a rotameric minimum might help drive the catalytic reaction.[19] Irene Luque and Ernesto Freire describe how the protein binding sites are often characterized by regions with high stability and regions with low stability.[14, 20] Catalytic residues in enzymes are usually in highly stable regions which may allow for preorganization of the binding site. Low stability regions have been shown to play an important role in allosteric enzymes, allowing for communication between an allosteric binding site and the active site. Some instable regions have been shown to be necessary for proper protein function.[21]

While there are some obvious differences between protein-protein binding hot spots and protein-ligand binding, there are some interesting parallels, such as the role of rigid and flexible residues at hot-spots. Nussinov showed that there are a few key rigid residues at protein-protein hot spots that act as anchors, surrounded by flexible residues.[22, 23]

1.3 Scoring Protein-Ligand Binding

A natural test of scientific understanding of protein-ligand binding is to be able to predict a binding mode and its affinity. Predicting the structure of the complex of a small ligand with a protein (molecular docking) is a complex task. Docking is divided into two parts, sampling and scoring. The first part is sampling of the conformational space of a protein-ligand complex, identifying modes of the ligand binding to the protein. The second part is evaluating and scoring the quality of interaction between the ligand and protein. Ideally, the scoring function would be able to estimate binding affinities of the bound pose or, at least, rank order the list of conformations. A majority of the effort in docking research has focused on improving the scoring functions, as it has become obvious that scoring functions have significant room for progress.[4, 3]

Scoring functions can be classified into three primary classes and a fourth class being a hybrid mixture of the first three classes. The first class is composed of force-field based scoring functions, which use the classical molecular mechanics force fields such as CHARMM and AMBER.[24, 25] For example, I.D. Kuntz, using an AMBER-based force field was able to dock, score, and identify the family of orientations closest to the experimental binding geometry.[26] Force-field based scoring functions include GOLD[27], AutoDock[28], and DOCK[29].

A second class is empirical scoring functions. Empirical scoring functions break the binding free energy into different types of interactions, such as hydrogen bonds, ionic interactions, hydrophobic contacts, and entropic penalties.[30] The functional forms in empirical scoring functions follow force-field based scoring functions, although they are often more simple.[31] A hurdle for developing empirical scoring functions is assigning appropriate weights for each of the terms. Large and diverse training sets with binding data are obviously important in order to generalize results to different systems (see Table 1.1). Examples of empirical scoring functions include Score1(LUDI)[32], Score2(LUDI)[33], F-Score[34], ChemScore[35], ProteusScore[35], HINT[36] and X-Score[37].

The third class comprises knowledge-based scoring functions. This class uses statistical analysis of structures to derive a sum of potentials of mean force between the ligand and the

protein. These are based on a statistical potential or knowledge-based approach, which is derived from pairing frequencies of protein-ligand atom pairs observed in a database such as the Protein Data Bank (PDB). In contrast to the empirical scoring functions, the knowledge-based scoring functions convert structural information scores without any knowledge of binding affinities, and thus may be more general because of a larger training set of available structures.[38] The theory behind the statistical potential approach is the assumption of a Boltzmann-distribution rule for a single, closed system held at fixed temperature that is applicable to a database of structures.[38] Examples of knowledge-based scoring functions include PMF[39], DrugScore[40], and SMoG[41], and ITScore[42, 38] (see Table 1.1).

The fourth type of scoring functions is "consensus" scoring functions.[31] There are various methods of combining scoring functions, such as voting (intersection), rank voting, weighted-sum ranks, and multivariate methods.[43] Examples include CScore by Tripos and DS Ligand Score by Accelrys.

Table 1.1 Size of protein-ligand datasets used to training scoring functions

	Year	Complexes
Score1	1994	54
F-Score	1996	19
VALIDATE	1996	65
ChemScore	1997	112
ProteusScore	1997	82
Score2	1998	94
PMF	1999	225
BLEEP	1999	90
DrugScore	2000	83
SMoG	2002	119
HINT	2002	53
X-Score	2002	230
ITScore	2006	851

1.4 Protein-Ligand Databases

Databases are essential for analyzing protein-ligand binding. Not only can they be used to develop and test scoring functions, but protein-ligand databases can be used for mining, to search for physicochemical properties that correlate with tight, specific ligand binding. The explosion of protein-ligand databases reflects the usefulness and interest in this area of

research.

AffinDB

AffinDB collects binding-affinity data for protein-ligand complexes in the PDB, in a "bottom up" approach.[44] It contains 474 PDB files with binding data for each PDB file. All of the entries have binding data collected primarily from literature and previously published collections of binding data.[37, 32, 45, 46] It does not explicitly exclude NMR or specify an X-Ray resolution threshold.

BindingDB

BindingDB, developed by Michael Gilson, contains a voluminous amount of very high-quality thermodynamic data.[47] It contains $\sim$ 20,000 binding data for $\sim$ 10,000 different ligands.[48] BindingDB now holds some K_i data in addition to its isothermal calorimetry (ITC) data. Additionally, experimental conditions are available for most of the data. BindingDB has increased the number of proteins covered to 1005, as of December 2008, considering any mutations as separate targets. Structural information is not available for many complexes.

CLiBE

Computed Ligand Binding Energy (CLiBE) is a database of computed binding energies for ligands in a set of PDB structures.[49] The binding energy is based on the AMBER molecular mechanics force field. This binding energy is calculated for each ligand. Where more than one ligand exists in a protein crystal structure, the energy for each ligand is calculated independently. Unfortunately, the definition of ligands is not clearly given, but might be assumed to be individual HETATM groups. Searching is based on PDB id, ligand name, or protein name.

DrugBank

DrugBank - a pharmaceutical database - contains extensive cheminformatic and structural information about drugs, and it contains bioinformatic and biological information about proteins (drug target).[50] However, DrugBank does not connect proteins and ligands together in a direct, structural fashion. For example, in the entry discussing Celecoxib, DrugBank

presents extensive cheminformatic information(Drug Interactions, Half Life, InChi, LogP, LogS, etc) and discusses two possible protein targets (one of which has a crystal structure, 2BIY, but this structure does not have a Celecoxib in the crystal structure). There is no specific information connecting the ligand to the proteins. Neither does it mention the PDB structure of carbonic anhydrase II structure crystallized with Celecoxib (1OQ5).

*e*F-Site

*e*F-Site is not strictly a protein-ligand database, but rather a database of protein pockets with ligands.[51] However, it does link protein with their ligands because they are used to define the binding pocket. *e*F-Site focuses on the electrostatic and physicochemical mapping of all sites on proteins, using 5 Å distance cutoff. The authors used their technique for functional annotation of a hypothetical protein to identify which ligands might bind to a given site, based on homologous structural information.[52, 53] The authors have since developed tools to allow searching and comparison among binding pockets.[54, 55]

FireDB

FireDB is a database of residues involved in ligand binding for a set of PDB structures, annotated with functionally important residues.[56] Important residues are identified from Catalytic Site Atlas (CSA) and consensus sequences generated from protein sequences clustered 97% sequence identity.[57] Protein residues are identified at distance cut offs of 3.5, 4.0, and 4.5 Å. To limit the ligands to biologically important small molecules, solvents, buffers, ions, DNA, RNA, peptides, very large ligands (where the ligand has two-thirds of the atoms of the protein) are excluded. FireDB is a source of important residues for their FireStar server, used to predict functionally important residues using SQUARE[58, 59]. FireDB strives to be inclusive including data from NMR and does not have a resolution limit for crystal structures.

HIC-Up

HIC-Up also lists all HETATM groups in PDB files, as well as links to other resources for further information.[60] HIC-Up is updated twice a year. HIC-Up provides a number of ways to search for HETATM groups, including QuickXS mechanism, HETATM code (three-character code), trivial name (e.g., benzene), chemical formula (e.g., C6 H12 O6),

number of non-hydrogen atoms index lists of chemical composition as well as by standard search engines (google, AltaVista).

HET-PDB

HET-PDB Navi is a database of HETATM groups in the PDB.[61] It includes all HETATM groups in the PDB file, as long as the HETATM group is not one of the basic twenty amino acids or water. No affinity information is available.

Iditis

Though Iditis is no longer available, it provided access to the PDB files in a relational database form.[62] This allowed direct Structured Query Language (SQL) queries against the data. A graphical interface was available in addition to the command line access. Several programs were used to extract the data from the PDB file and populate key tables. For NMR files, various programs were used to select a few representative models from the ensemble. Data available for queries included file properties (name, function, authors, dates, resolution, etc), sequence, secondary structure, hydrogen-bonding interactions, electrostatic interactions, torsion angles, solvent accessibility, and ligand groups.

KDBI

KDBI (Kinetic Data of Bio-molecular Interactions) is a collection of kinetic data for macromolecular interactions, whether they are protein-protein, protein-ligand, DNA-ligand, or RNA-ligand interactions.[63] While there is information regarding proteins and ligands, the focus is on pathway information. There is no structural information in the database.

K*i*Bank

K*i*Bank strives to connect K_i data to protein structural data, with the goal of providing QSAR data sets.[64] KiBank first collected binding data taken from literature abstracts and generated a 3D ligand structure from a 2D structure. KiBank then chose a crystal structure from the PDB as representative for the protein (the crystal structure does not need to have the ligand crystallized with the protein). Lastly, it added hydrogens and minimized the structure. Though papers from the group have appeared as recently as 2006, it seems that KiBank is no longer available.

LigASite

LigASite, a database of binding sites, has the motto, "A gold-standard dataset of biologically relevant binding sites in protein structures".[65] It excludes NMR structures and has X-Ray resolution limit of 2.4 Å. Like other protein binding-pocket databases, it uses ligands to define binding sites, but restricts ligands to those HET groups that have at least 10 heavy atoms, and defines the pocket as those residues within 4 Å of the ligand. However, this database does not provide an easy mechanism to identify what ligands are used to define the binding pocket for each protein. What is novel about LigASite is how the binding sites are represented. The binding pockets are presented by apo proteins and the ligand site definition is defined in a crystal structure of the same protein with a ligand (or more than one). The matching of holo-protein structures with apo-protein structures should allow for interesting studies for ligand binding induced changes.

Ligand-Protein DataBase

The Ligand-Protein DataBase (LPDB), developed by Charles L. Brooks III in 2001, has roots in improving empirical scoring functions.[30] LPDB contains 195 complexes corresponding to 51 different receptors (21 protein classes), annotated with binding data, with 178 unique ligands. The data are collected from six empirical scoring function training sets. The LPDB is designed to be used along with a continuum set of pre-generated decoys to assess scoring-function performance.

MSDsite

MSDsite is designed to allow extensive querying of the binding sites in the MSD database.[66] This database contains all PDB ligands, their interactions with macromolecules (protein, DNA, and RNA), coordination, protein sequences, and ProSite motifs. It aims to annotate ligands with their topology, bond orders, and hybridization by means of their database MSDchem. Annotated interactions including covalent bonds, ionic, hydrogen bonding, van der Waals, planar groups (for certain groups of four or more atoms that are recognized as being planar with each other), and "non-bonding interaction" for atom pairs within 4 Å of each other but no bond type has been classified. Ligands are defined as any HETATM group, including modified amino acids that are part of the main protein chain. MSDsite allows for searching for any protein in the MSD database or for any PDB file uploaded by a user. This database provides means to filter by experimental method (NMR,

X-Ray, Electron Microscopy, Theoretical), X-ray resolution, ProSite pattern, ligand and protein content, as well as macromolecular target (DNA, RNA, or protein).

MSDMotif

MSDmotif is a database which aims to provide a mechanism for querying the PDB about motifs.[67] It provides an extensive web interface for provides searches by PDB header file information, small-molecule information, small 3D motifs (e.g. beta-bulge, beta-turn, schellmann-loop, st-turn, etc), secondary structure elements, protein sequence patterns, as well as protein side-chain, main-chain bonds and protein-ligand interactions (covalent, ionic, hydrogen bonds, van der Waals, π interactions, and unidentified interactions within 4 Å. It provides the search results in a number of formats. The software to search for the annotations is also provided under an Open Source software license. Ligands identification is based on MSDsite. There is no protein cleanup.

PDBbind

Currently, PDBbind contains binding data on 3214 complexes, with 2084 unique ligands, collected from the PDB.[68, 69] PDBbind was curated in a very similar fashion as Binding MOAD but has some key differences. PDBbind focuses on complexes with only one ligand in the crystal structure. PDBbind also excludes any complex binding a simple cofactor such as ATP. While Binding MOAD uses a resolution threshold on crystal structures of 2.5 Å, PDBbind has no threshold value (the largest crystal structure resolution is 4.7 Å). Binding MOAD also provides information on the structures when we do not have binding data because they are still a valuable resource in database mining, and for knowledge-based scoring functions while PDBbind only provides structures of complexes for which it has binding data. The research projects around PDBbind focus on developing scoring functions and searching ligand substructures.[70, 71, 72]

PDBcal

PDBcal focuses on ITC data for receptors (both proteins and nucleic acids) with structures in the PDB[73]. For a given receptor, PDBcal attempts to identify the most relevant structure. The PDB file referenced may have the ligand in the crystal structure, but it is not required (e.g. for the protein concanavalin A with ligand pyranoside 1 the PDB file 1GKB is referenced but contains no ligands). The referenced PDB structure may be an NMR structure,

and no threshold resolution on X-Ray structures is used. Furthermore, there are examples of a given protein, with data for different ligands, and each references the same PDB structure. PDBcal makes the point that all data is extracted from the literature and not collected from other existing datasets.

PDBLIG

PDBLIG was a database that attempted to match protein domains to ligand binding.[74] Protein domains were classified using the CATH protein classification system. Intensive effort was spent in appropriate definition of ligands, as ligands were analyzed for attachment to other ligands and interactions with the protein. Separate HETATM groups were linked if they were within covalent distance of each other (where distance cutoffs were used for different atom types). Peptides were counted as ligands if shorter than 30 residues long. Covalently attached ligands were separated from the protein and treated as valid ligands. Ligands with missing coordinates or density were left as found in the PDB file. Bond orders and formal charges were assigned to the ligands using HBADD. Interactions between ligand and protein were calculated with LIGPLOT and HBPLUS. Ligands that were found in between protein units or had contact with fewer than five residues were discarded. Additionally, ligands were classified into categories such as cofactor, nucleotide, sugar, organic or peptide. Properties of the ligands were calculated to measure compliance with Lipinski's "rule of five".[75] This database aimed at asking if different ligand types associated with certain protein families and if protein families bound functionally or structurally similar ligands. This database was initially generated by Inpharmatica, but this company was acquired by Galapagos and PDBLIG no longer exists.

PDB-Ligand

PDB-Ligand was designed as a tool to structurally align all of the ligand-binding pockets in the PDB, based on a flexible ligand-clustering method.[76] The clustering uses the RMSD value between all residues within 6.5 Å of the ligand after alignment via the Kabsch method.[77] The authors cluster 161 PDB files (which have 321 ATP binding sites) and cluster them into 165 different clusters using 0.5 Å RMSD cutoff, and 91 clusters using a 1 Å cutoff.

The database contains NMR proteins and does not have an X-Ray threshold for the proteins. It also contains nucleic acid structures. All HET groups are defined as ligands, even if they are covalently attached to the protein or are simple ions such as magnesium or

sodium. The website allows searches based on ligands, but it is case-sensitive. There is no straightforward way to search based on proteins. A helpful statistics page citing some key statistics is available. http://www.idrtech.com/PDB-Ligand/

PLD

The Protein Ligand Database (PLD) by John Mitchell is a small database of protein-ligand complexes.[78] All of the entries are annotated with calculated binding energy using the knowledge-based method BLEEP. Of the 485 entries, 345 were annotated with experimental binding data. While ligand similarity scores have been calculated, they are not available; additionally, the search functionality does not work. http://www-mitchell.ch.cam.ac.uk/pld/index.html

ProNIT

The Protein-Nucleic Acid Interactions (ProNIT) database provides experimentally determined interaction data between proteins and ligands that are nucleic acids.[79, 80] It contains information on the proteins and nucleotides as well as binding data (e.g. K_d, δG, δH, δC_p) as well as the bibliographic source of the information. No information as to the maximum or minimum length of the nucleic acid ligands. (It may well be that there are no single-nucleotide ligands in this database).

Relibase

Relibase (*Re*ceptor *Li*gand Data*base*)was originally developed by Manfred Hendlich in 1998, is currently maintained by Cambridge Crystallographic Data Centre (CCDC) in collaboration with Gerhard Klebe.[81] Relibase has developed into commercial application, called Relibase+, using both PDB and proprietary databases. Relibase+ touts an extensive collection of tools for searching and data mining. Tools include developmental tools such a command-line interface and python libraries, as well as graphical user interface (GUI) tools to explore topological similarities of cavities, crystal-packing effects, hotspot interactions, water-mediated interactions, and protein-ligand binding.[82, 83, 84, 85]

Relibase is a subset of Relibase+, and is available upon registration to academic users. However, almost all of the tools are limited to the commercial version. Relibase+ uses a very broad definition for ligands. Relibase+ considers almost all HETATM groups in crystal structures as ligands. Ligands include metal cations like magnesium, inorganic salts such as

sulfate, common crystal additives like polyethylene glycol, and even modified amino acids within the protein chain.

sc-PDB

The screening-PDB (sc-PDB) database was designed for database and inverse screening of a subset of binding pockets in the PDB.[86, 87, 88] The authors created similar criteria to those in Binding MOAD, limiting ligands to biologically relevant ligands and requiring an X-ray resolution threshold of 2.5 Å. They divide their ligands into nucleotides, peptides, small organics, and cofactors and show that the data is widely diverse.[87]

In the first screening study, the authors used GOLD to recover the known targets for four unrelated ligands in an inverse screen. However, they cautioned that the screening procedure be applied only to selective compounds after finding much less accurate inverse screening results with adenosine monophosphate (AMP).[86] In two different screening studies, the authors were able to show 70- to 100-fold enrichment of ligands and that inverse screening is possible even with some revealed limitations.[86, 88]

SitesBase

SitesBase, developed by Nicola D. Gold and Richard M. Jackson from the University of Leeds, is a database of protein-ligand binding pockets.[89, 90] Gold and Jackson apply two different scoring functions to calculate the similarity of binding pockets. The first technique is the "seq sim" score, which is calculated according to the method of Stark and Russel.[91] The second method is the Poisson Index scoring scheme. The Poisson Index is calculated with the number of size and matching atoms between two binding sites.

Their intention is to identify how different protein binding sites maintain selectivity and specificity for their ligands To facilitate this, they predict cross-reactivity (i.e., predict drug side effects) and provide functional annotation of new and existing proteins.[92] In fact, the authors were able to show cases where two proteins are diverse in structure, function and sequence and yet share a common binding site. For example, SitesBase finds Subtilisin DY (1BH6) is similar to proteinase A (1SGC). The two proteins do not share any significant sequence similarity. However, their algorithm found that the two ligands, a synthetic inhibitor N-benzyloxycarbonyl-ALA-PRO-PHE-chloromethyl ketone (1BH6) and chymostatin (1SGC), are extremely similar in their score and can be superimpositioned very well.[92]

The authors do not exclude NMR structures, nor do they have a threshold resolution on

X-ray structures. The authors do exclude ligands that have less than six atoms, some ligands that are not biologically relevant (e.g., TRIS buffer), peptides, and post-translationally modified residues. However, SitesBase has problems with ligands that are composed of more than one HETATM group in the crystal structure as it treats each HETATM group as separate ligands, even if they are covalently attached to another HETATM. This database is not browseable, downloadable, and not available to commercial entities. Furthermore, it has not been updated since its initial release in 2006 using data available from the PDB as of June 2005.

SMID

Small Molecule Interaction Database (SMID) was part of the Blueprint Initiative, and provided predictions of protein domain-small molecule interaction for proteins in Biomolecular Interaction Network Database (BIND).[93, 94] SMID clustered known cases of similar protein-ligand binding in attempts to find binding patterns. SMID used a tool, SMID-BLAST, to predict a domain binding site from a given input sequence, from homologous protein sequences. SMID-BLAST then provides a list of potential small molecule ligands based on SMID scores and aligned binding pockets. SMID allowed searches based on ligands or protein sequence. Unfortunately, it was unclear how ligands were defined from the PDB structures in BIND.

SMID-BLAST was validated by trying to predict the ligand for 793 proteins crystal structures, of which 472 (60%) matched the observed small molecule in the crystal structure. However, the Blueprint Initiative failed and BIND and SMID are no longer freely available. They have been sold to a commercial entitity.

SuperLigands

SuperLigands is a database that focuses on the ligands in the PDB.[95] It provides searches by 2D chemical similarity to each other and to a set of 2396 drugs, 3D superimposition, and provides the ligands in the MDL Mol file format which, in contrast to the PDB format, includes information about bond types. Some cheminformatic information is available for the ligands, such as how well each ligand measures against Lipinski's "rule of five".[75] Protein information is limited to the PDB id for the protein-ligand complex.

Structure and Cheminformatic Respositories

As the Protein Data Bank grew and developed, national laboratories such as the National Center for Biotechnology Information (NCBI) and European Bioinformatics Institute (EBI) have mirrored the data in the PDB and provided their own interface to the data. NCBI developed the Molecular Modeling DataBase (MMDB) and EBI developed PDBsum and E-MSD. Furthermore, each of these institutions has developed its own cheminformatic databases and interfaces, which allow mining for ligands. Research Collaboratory for Structural Bioinformatics (RCSB) developed Ligand Depot (now called the Ligand Expo), NCBI has PubChem, and EBI has MSDchem.

1.5 Conclusion

In this dissertation, I will describe how I have developed Binding MOAD and then mined it for principles of protein-ligand binding, examining protein flexibility and ligand binding.

Chapter 2 discusses how Binding MOAD was developed. It describes the criteria used to select crystal structures from the PDB for inclusion in Binding MOAD. It discusses the extraction of binding data for the protein-ligand complexes from the literature. This chapter describes how the data are organized for useful presentation. Finally, it gives a summary description of the data in Binding MOAD.

Chapter 3 discusses a data-mining study of Binding MOAD. Because of the careful and extensive annotation of binding data that went into Binding MOAD, it is a very appropriate dataset for analyzing physicochemical properties that correlate with specific, tight binding. This chapter discussed how mining revealed differences between enzymes and non-enzymes, and the details of those differences, along with possible consequences with regard to drug design. Namely, this chapter shows how careful examination of the data reveals how divergent approaches may be more productive for improving the affinity of ligands for the two types of proteins. It also reveals inherent flexibility differences in the amino-acid content of enzymes versus nonenzyme binding sites.

Chapter 4 describes how Binding MOAD was used to survey protein flexibility in crystal structures. While crystal structures are widely used as static protein models in SBDD, protein flexibility needs to be accounted for to accurately predict ligand binding. This chapter measures ligand-induced changes (both in backbone RMSD and side-chain flexibility) as seen in protein-ligand crystal structures. Both intrinsic and extrinsic variation in flexibility from complexed and uncomplexed structures were investigated.

Appendix A discusses the architecture of the server used to distribute Binding MOAD

online. It discusses how Binding MOAD uses a 3-tier model (client, application, and database) implemented with a JBoss Application Server using the Jakarta Struts MVC framework and a MySQL database for persistence. The appendix discusses how and why these tools and technology are used to build the webserver for http://bindingmoad.org.

Appendix B discusses development of Natural Language Processing (NLP) tools for use in the annual updating of Binding MOAD. The manual extraction of of binding affinity data from the literature is a bottleneck step in the annual updating process. By turning to NLP, a field of artificial intelligence and linguistics that turns human language into a form that can be processed by computers, the step of extracting binding data from literature is partially automated. This partial automation significantly speeds the annual updating process.

In conclusion, this dissertation will show that Binding MOAD was carefully crafted and is aptly able to illuminate our understanding of protein-ligand binding.

Chapter 2

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[32] and more recent sets are 200-800.[30, 96, 68] 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)[61] 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 11,368 structures, with binding data available for 3453 (30%) 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.[30] 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.[30]

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).[96] Binding DB has recently started to accept the deposition of K_i data, and the number of entries has grown significantly to 3300 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.[68] It contains binding data on 800 complexes with resolution 2.5 Å (559 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.[97, 98] 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.[83, 99] MSDsite is the newest resource in the MSD suite of web-based tools from the European Bioinformatics Institute.[66] 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[100, 101, 102] 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.[103] 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 11,368 complexes in Binding MOAD, there are 3582 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 3582 complexes in the nonredundant set, we have obtained binding data for 1008. (In cases were 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.[104, 86]

2.2 Methods

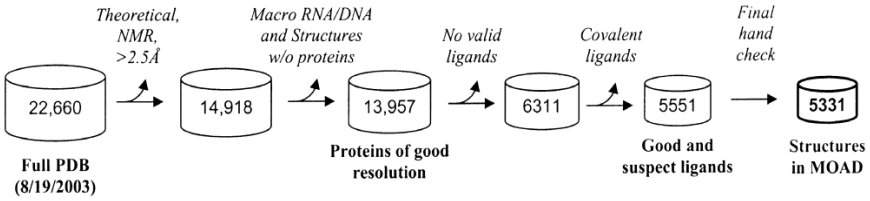
2.2.1 Top-Down Approach

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,[61] 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[105] from the Research Collaboratory for Structural Bioinformatics (RCSB) and

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.



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.[106] 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.[86] 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.

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-

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.

containing enzymes where a zinc-ligand distance < 2.1 Å is not necessarily a covalent bond.[107] 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 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.

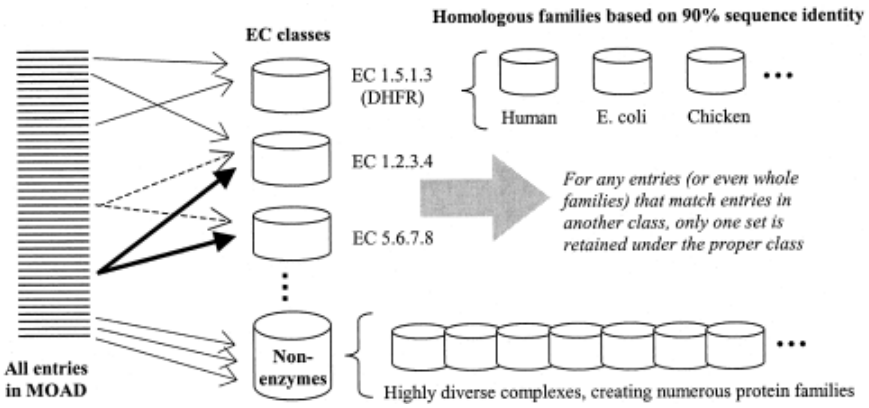
Modifications to amino acids are on the partial ligand list because they can be part of the protein or part of a peptide ligand. Complexes containing heme groups were rejected because the covalent association of ligands to the central metals made it difficult for us to properly identify the true ligands. 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.[108] 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.[109] 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,[110] but our grouping of proteins is slightly different than the clustering used for grouping similar sequences at the PDB.[111] The routine is presented in Figure 2.2:

Figure 2.2 Currently, 3582 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

After examining the PDB contents in our latest updated, January 1st, 2008 (48,178 entries), a total of 11,368 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[86] due to slightly different selection criteria. Three-fourths of the proteins are enzymes, with hydrolases and transferases having the most representatives.

Table 2.2 Functional classification of current entries in Binding MOAD

Proteins identified with EC numbers[a]	**Entries**[b]
1.-.-.- (OXIDOREDUCTASE)	1914 (16.8%)
2.-.-.- (TRANSFERASE)	2495 (21.9%)
3.-.-.- (HYDROLASE)	3155 (27.8%)
4.-.-.- (LYASE)	653 (5.7%)
5.-.-.- (ISOMERASE)	427 (3.8%)
6.-.-.- (LIGASE)	254 (2.2%)
Total enzymes	8927 (78.5%)

Proteins without EC numbers	**Entries**
Binding (lectin, streptavidin, agglutinins, etc.)	537 (4.7%)
Signalling, cell cycle, apoptosis	376 (3.3%)
Folding (chaperones, etc.)	55 (0.5%)
Immune (antibodies, immunoglobulins, cytokines, etc.)	254 (2.2%)
Mobility/structural (actin, myosin, etc.)	79 (0.7%)
Toxin/Viral	81 (0.7%)
Transcription, translation, replication proteins	263 (2.3%)
Transport (amino acid transporters, electron transport, etc.)	382 (3.3%)
Enzymes without EC numbers (eg., isopenicillin N synthase)	65 (0.6%)
Other	349 (3.1%)
Total proteins without EC numbers	2441 (21.5%)

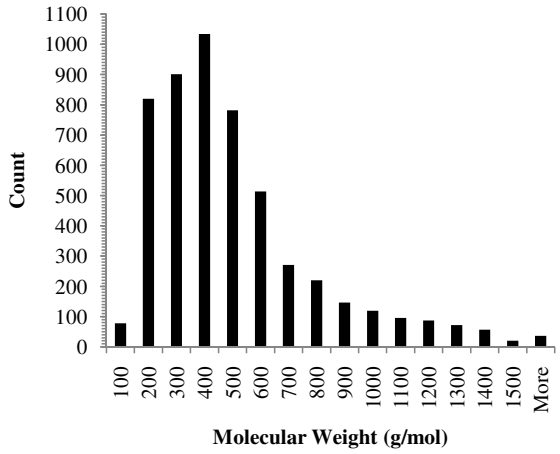
[a]Enzyme counts include entries without EC numbers that could be identified through keywords or enzyme names. Some were also identified by 90% sequence identity to entries with EC numbers.

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

Binding MOAD contains 5358 unique, valid ligands within the 11,368 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.

Figure 2.3 Distribution of the current 5358 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.



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, 6321 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 protein families increases (Table 2.3). Clustering at 90% sequence identity (our preference) produces 3582 homologous protein families with the largest family containing 246 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

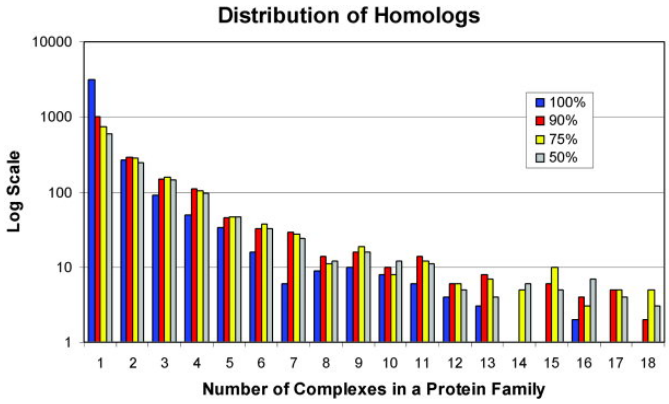
various ligands have been co-crystallized.

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	6321	124 complexes[1] (52)[2]
90% Sequence identity	3582	246 complexes[3] (94)[1]
75% Sequence identity	3305	178 complexes[3] (94)[1]
50% Sequence identity	2889	186 complexes[3] (111)[1]

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

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).



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, 1857

of the 3582 families contained only one complex, and so the choice for the representative was obvious. The remaining families contained multiple complexes. For 724 of the families, the representative was easily identified by binding data. Resolution was the deciding 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.)

The nonredundant version of Binding MOAD contains 3582 unique proteins. After choosing the complexes for the nonredundant set as outlined above, this set contains binding data for 1013 of the unique structures.

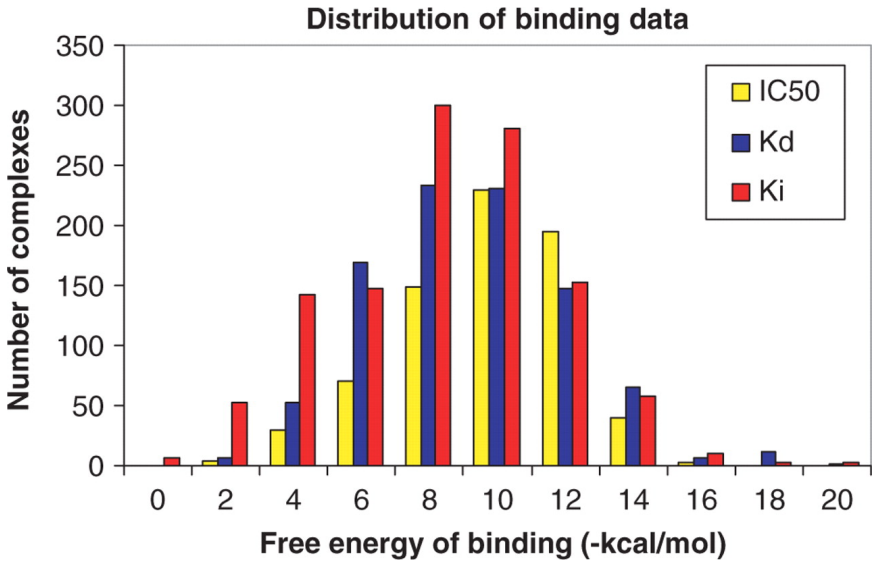
2.3.3 Binding-Affinity Data

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. Only 159 entries have IC_{50} data, ranging 60 pM - 14 mM. For the 516 entries with K_d data, values range 190 fM - 250 mM. The 700 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 11,368 complexes). At this time, only 3453 complexes (30%) in Binding MOAD are augmented with binding data. Though this is much larger than other datasets with a few hundred binding affinities,[30, 96, 32] 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.[68]

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.[112, 113]

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 and 11,368 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 crystal 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[114]. 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.

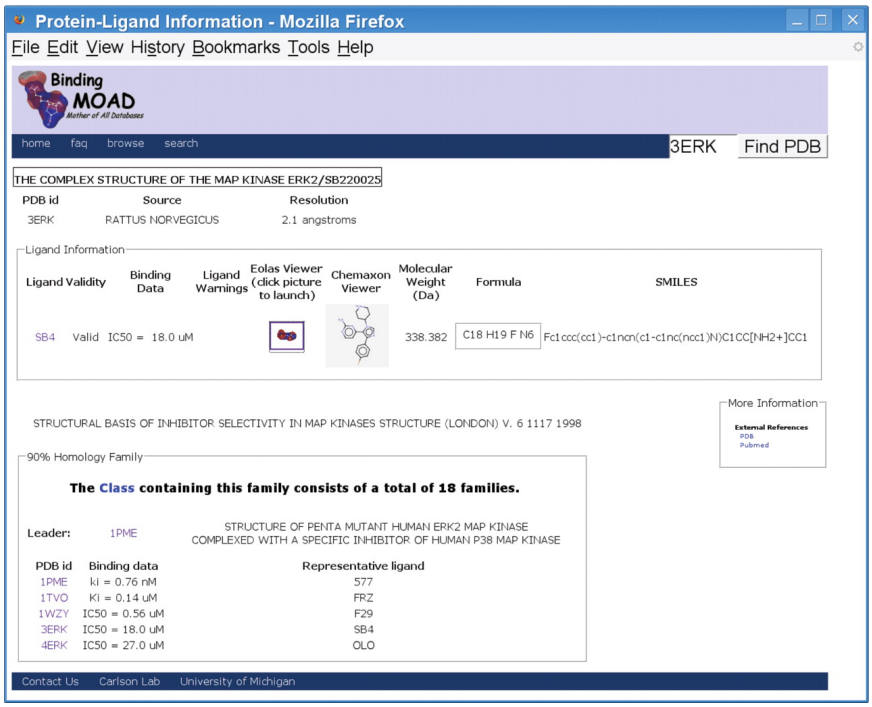
Platform

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.

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.



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.

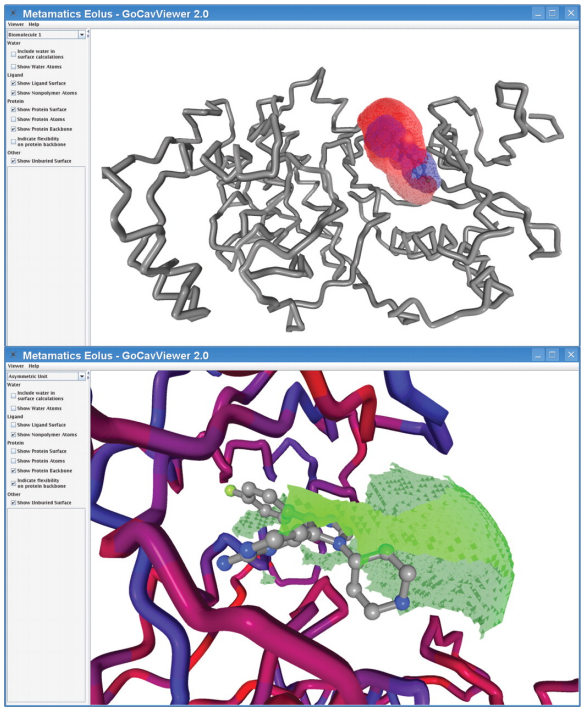
By taking advantage of rendering algorithms and OpenGL Shader Language (GLSL), Eolus provides the new viewer with new representation styles. The surface representation has been expanded to a fully transparent polygon surface. The proteins are rendered as ribbons by default, and the entire protein (instead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

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

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.



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://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 3

Differences Between High- and Low-Affinity Complexes of Enzymes and Nonenzymes

3.1 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.[115, 116]

Binding MOAD is the largest curated database of high-resolution protein-ligand complexes from the Protein Data Bank (PDB).[61] 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.[75, 117] 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.[118] 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.[119, 120] 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[121, 122] 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 where 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.[123, 124, 125, 126] In this study, ligand efficiencies for the different classes of proteins are reported as affinity per size ($-\Delta G_{bind}$ divided by the number of non-hydrogen atoms) and per the degree of contact between the ligand and the pocket ($-\Delta G_{bind}$/BSA).

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

3.2 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 12/31/2006 or earlier; it contains 9836 complexes, comprised of 3151 unique protein families binding 4659 unique ligands. The great care taken in curating this dataset has been outlined elsewhere,[115] but it should be noted for these purposes that ∼9,000 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 $\Delta G_{bind} = -RT \cdot \ln(K_d)$ or simply approximated by $\Delta G_{bind} = -RT \cdot \ln(K_i \text{ or } IC_{50})$ with a temperature of 298 K.

High-affinity binding was defined K_d, K_i, or IC_{50} $\leq$250 nM (ΔG_{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. 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.[116] 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,[127] based on the method developed by Wildman and Crippen.[128] 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.

3.2.1 Statistical Analysis.

Statistical significance was assessed with the programs SAS[129] and JMP[130]. 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.

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.

3.3 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[131] based on the PDBbind dataset[69] 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,[123] we changed our approach and focused on distributions of the properties between subsets of protein-ligand complexes.

Table 3.1 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 3.1 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 $\sim$300 Å^2 of BSA, but the high-affinity complexes display more contact. It has been estimated that drug-like binding sites have $\sim$300 Å^2 of solvent-accessible surface area (SASA).[121] Our measurement for BSA is based on MSA, and so, the slightly higher values of the high-affinity complexes are appropriately comparable.[121]

3.3.1 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 3.1). 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 3.1, Figure 3.1b). 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.

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 3.1). Distributions of the calculated octanol/water partition ratios (Figure 3.1a,b) show that high-affinity ligands are more hydrophobic than those with

Table 3.1 Characteristics of Protein-Ligand Binding for Enzymes and Non-Enzymes in the Full Dataset.

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 (%ESA)[d]	87 Å^2 (22%)	144 Å^2 (24%)	
SlogP	0.3	2.4	
-ΔGbind/atom	0.31 kcal/mol-atom	0.36 kcal/mol-atom	
-ΔGbind/BSA	21 cal/mol-Å^2	26 cal/mol-Å^2	
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 (%ESA)[d]	118 Å^2 (33%)	45 Å^2 (11%)	
SlogP	-2.2	1.5	
-ΔGbind/atom	0.28 kcal/mol-atom	0.41 kcal/mol-atom	
-ΔGbind/BSA	22 cal/mol-Å^2	31 cal/mol-Å^2	
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 3.1 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).

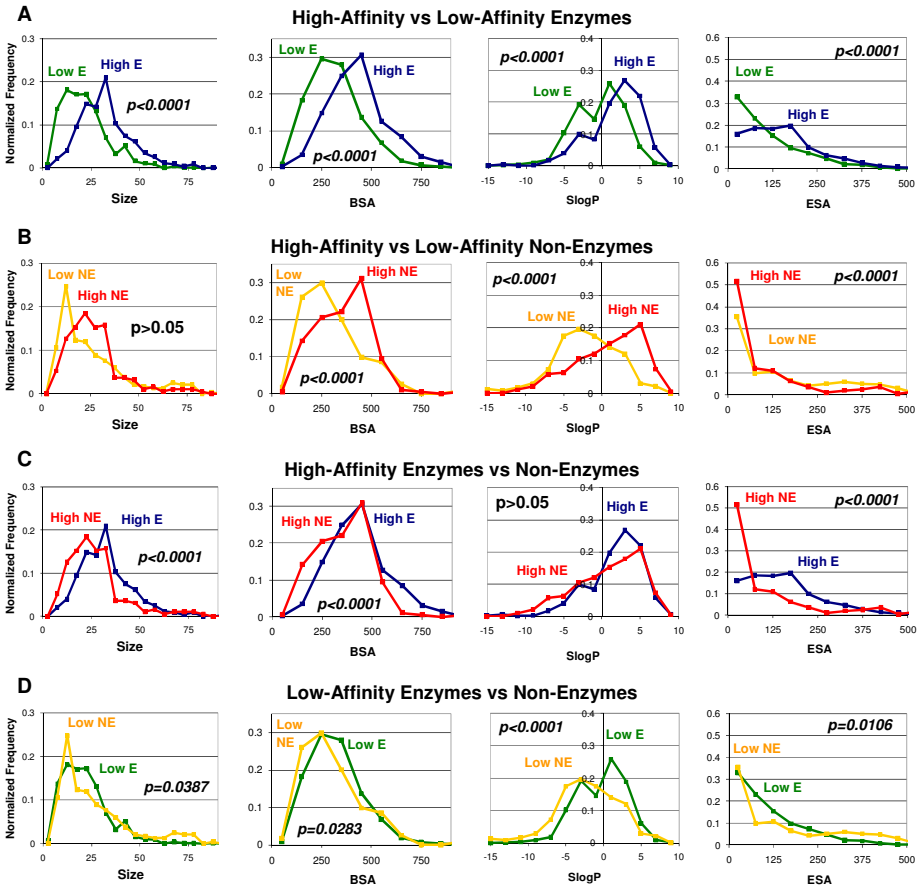
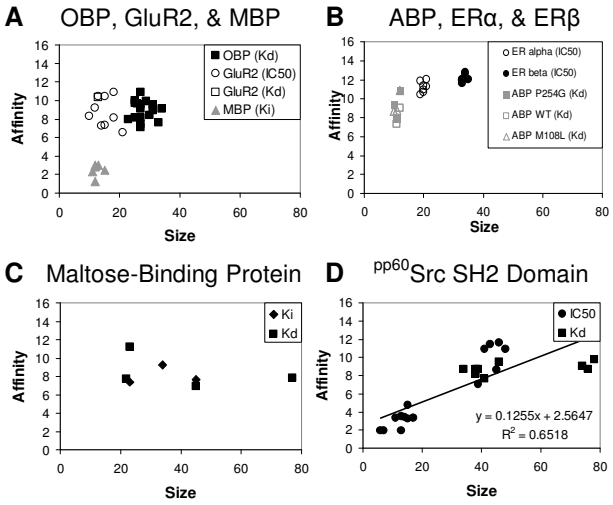


Figure 3.2 Limited correlation is seen between size and affinity in non-enzymes (A and B). The proteins with "clusters" of points have smaller binding sites and no ligands over 40 non-hydrogen atoms. The ligands have similar sizes and affinities for oligopeptide-binding protein (OBP), glutamate receptor 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).

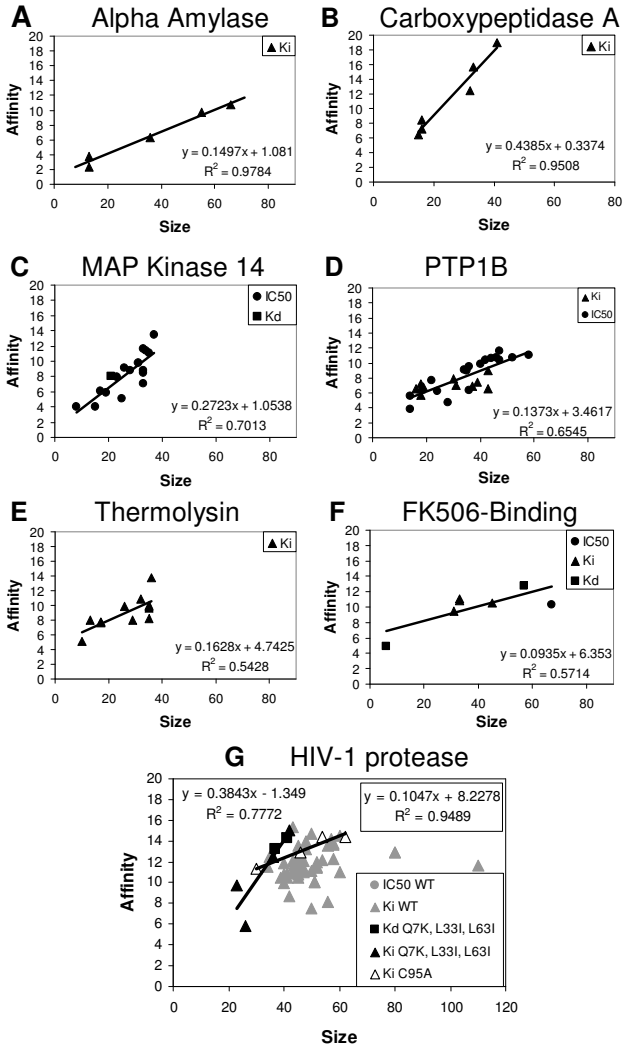


low affinity, but there is no significant difference between enzymes and non-enzymes in this regard. It appears that "adding grease" equally improves binding to both enzymes and non-enzymes, consistent with a general desolvation effect.[119]

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 3.3 and 3.3), 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:

Figure 3.3 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.



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 3.3 a, 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.

Only maltose-binding protein (Figure 3.3c) and the non-enzymatic site on the SH2 domain of ^{pp60}src tyrosine kinase (Figure 3.3d) have a significant range of ligand sizes. The maltose-binding protein complexes contain sugar chains of varying length. 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 3.3. One of the most interesting features of the data in Figure 3.3 is that the slopes - the overall trend for each set - vary significantly! 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 3.3b) to 0.09 kcal/mol-atom for FK506-binding protein (Figure 3.3f). 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.[72]

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

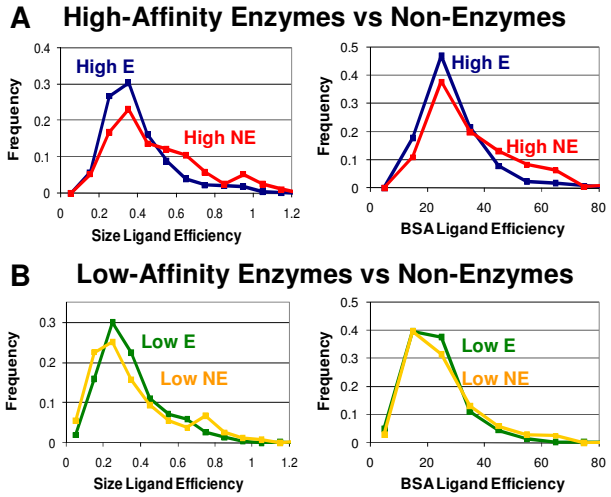
range of ligand sizes and a large range in binding affinity. It is debatable whether these trends are exceptional examples of the correlation expected for enzymes or whether they indicate cases where only conservative changes in sizes are allowed, as would be expected for non-enzymes. It is also possible 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,[132] fragment-based design often couples these small building blocks in the pursuit of high-affinity ligands.[133] 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.[134] 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 3.1.

Hajduk's trends were presented for the most efficient ligands for each protein, emphasizing the most ideal cases of improving a ligand.[134] 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 3.3g) 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.[134] 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.[135] 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

Figure 3.4 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.



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 3.1 and 3.3. 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.

3.3.2 Ligand Efficiencies.

Distributions of ligand efficiencies are given in Figure 3.4. 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.[136, 137, 138] 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.[138] 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 3.1, Figure 3.4b). 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 3.1). Non-enzymes in Figure 3.4a show a broader distribution of efficiencies and much higher populations above 0.4 kcal/mol-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. 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%).

3.3.3 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.[139, 122, 140, 141] 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.[142, 143] Recent discussions on the druggability

of protein-protein interfaces note that these difficult targets may be more amenable than originally thought.[144, 145] 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,[140] and genomic analysis has indicated many more receptors are druggable.[146]

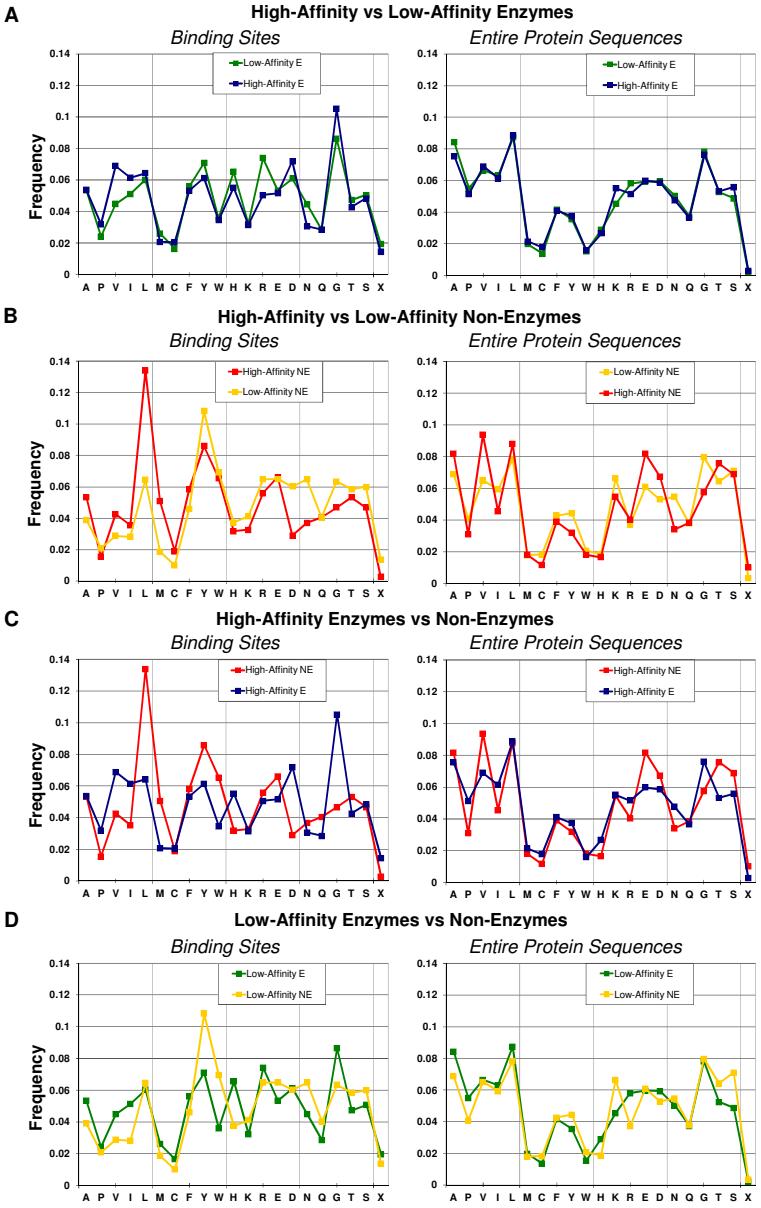
Our results are also in good agreement with a recent study that estimated the druggability of 1096 non-redundant human proteins.[122] The predictions used a statistical model trained on NMR-screening data using a small fragment library.[147] Four of the top six classes were non-enzymes: vitamin-binding, steroid-binding, lipid-binding, and nucleotide-binding proteins.[122] 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.

3.3.4 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 3.1c). *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 3.5, 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

Figure 3.5 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.

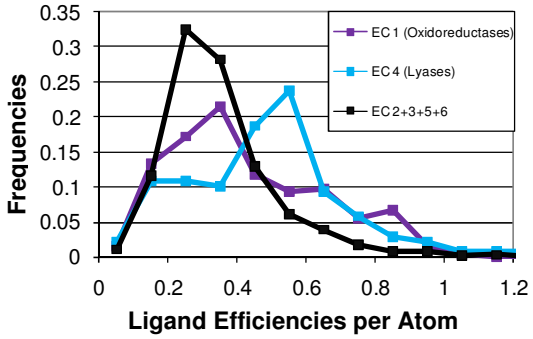


(Ala through Trp) on Figure 3.5 are 47.0% of the binding sites for high-affinity complexes, but 43.9% for low-affinity ones. This is fitting with the aforementioned finding that the high-affinity ligands are slightly more hydrophobic. The comparison between binding sites of high- and low-affinity non-enzyme 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 3.5. 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.[148, 149, 150]

3.3.5 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[140] and a large percentage of new targets identified through genomic analysis.[146] The distribution of ligand efficiencies for the enzyme classes suggests that lyases and oxidoreductases are the most druggable enzymes, Figure 3.6. 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

Figure 3.6 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 ($p<0.0001$)



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.[122]

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.[121] 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.[151] 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 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.

3.4 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.[124] 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,[151, 134, 122] 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.[152, 153, 154, 155] 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 4

Protein Flexibility and Ligand Binding

4.1 Introduction

Proteins are naturally flexible biopolymers composed of a string of amino acids folded into a largely non-covalent structure.[156] This flexibility is often tightly coupled to its function, especially for enzymes. Understanding the flexibility in proteins is an important aspect in areas such as protein folding, protein engineering, and rational drug design.

A key feature of protein-ligand binding sites is that they have both characteristically rigid and flexible residues.[14, 157] Rigidity can aid in specificity and tightness of ligand binding. Flexibility allows for the ligand to enter the active site and can be involved in communication between allosteric sites and binding sites. Often clusters of residues near binding sites tend to be in strained conformations.[18, 19] Ligand binding was seen to induce strain in these residues, and it was hypothesized that this increase in internal energy could be used by the protein for catalysis and ejecting the ligand from the active site.

Being able to fully account for induced changes is especially important in protein-ligand docking. Protein-ligand docking is used to predict the orientation and direction of a ligand binding to a protein. While simple in theory, this task proves to be very difficult in practice.[3, 4] A particular issue, known as the cross-docking problem, is illustrative of the difficulties of accounting for protein flexibility in ligand binding. Cross docking attempts to dock a ligand from one crystal structure into the binding site of another structure of the same protein, but research shows that many ligands do not fit unless the protein is allowed to adjust to the ligand.[158, 159, 160, 161] The larger the required adjustment, the harder it is to accurately predict protein-ligand binding.[162] Protein flexibility needs to be taken into account to accurately represent protein-ligand binding.

There have been many studies examining the extent and properties of ligand binding

by comparing protein crystal structures with ligands (holo) and without ligands (apo). A number of studies have examined the local characteristics of their respective binding sites, such as side-chain flexibility, while other studies have examined global protein changes upon ligand binding. Most studies fell into two categories: backbone root mean square deviation (RMSD) analysis and side-chain rotation analysis.

Structural variation is smaller when assessed through backbone motion. Three studies used small sets of 8-20 proteins to analyze backbone RMSD. Gutteridge and Thornton found that enzymes bound to either a substrate or product tended to be more structurally similar to each other than to free enzyme (substrate and product structures had an average C_α RMSD of 0.36 Å while apo enzymes averaged 0.75 Å RMSD to the substrate structures and 0.69 Å RMSD to the product structure). Fradera *et al* found that the binding site's structure is preserved upon ligand binding as evidenced by the fact that the average all-atom, active-site RMSD changes $\leq$ 1 Å, that more than 90% of atoms in contact with the ligand move less than 1 Å, and that most binding sites had only modest changes in their electrostatic potentials.[163] However, they found that these small movements induced significant changes in volume and shape so that geometric similarity indices (η) ranged from 0.44 to 0.90. Finally, Heringa and Argos described how ligand binding was sufficient to induced strain and push some binding-site side chains into rotamers outside of the typical minima.[18, 19]

Gutteridge and Thornton followed their work noted above by looking for conformational changes upon ligand binding in a larger set of structures. Of 60 enzymes, $\sim$75% of holo-apo pairs had C_α RMSD of $\leq$1 Å. This C_α RMSD was contrasted with the C_α RMSD observed among apo-apo protein pairs as a baseline, where $\sim$83% of 31 apo-apo pairs had a C_α RMSD of $\leq$1 Å.[164] Catalytic residues were observed to have more rigid backbones compared to other residues in the active site, as measured by C_α RMSD. However, the difference in rigidity was limited to the backbone; catalytic residues had equally flexible side chains as noncatalytic residues.

Gunasekaran and Nussinov classified 98 proteins into three categories based on maximum C_α displacement between holo and apo structures: rigid proteins ($\leq$ 0.5 Å), moderate (0.5 Å $<$ and $\leq$ 2.0 Å), and flexible ($>$ 2 Å).[1] All classes had the same contact density, so flexibility in certain residues was not due to loose packing. Rigid and moderately flexible proteins were seen to have more polar-polar interactions: 35% and 34% for rigid and moderately flexible versus 28% for flexible proteins. Overall, most of the φ, ψ changes between apo and holo were minimal. All classes had a few binding-site residues with φ, ψ angles in poor regions of the Ramachandran map. There were more in apo than holo structures, and they tended to cluster near the binding site. (See Table 4.1).

Table 4.1 Percent of residues with backbone φ, ψ angles in disallowed regions of a Ramachandran plot. Data is shown for both the binding site and the entire protein. Data is taken from [1].

	Holo	Apo
Binding Site	1.2%	1.7%
Whole Protein	0.8%	0.9%

Brylinski and Skolnick evidenced that most apo-holo protein pairs did not exhibit a significant structural difference, and that holo-holo protein pairs exhibited even less change, using the C_α RMSD metric.[165] For 521 single domain apo-holo structural pairs, 80% had an RMSD $\leq$ 1 Å, and among a set of single domained holo-holo proteins, $\sim$92% had an RMSD $\leq$ 1 Å.

It is important to note that analyses of side chain reveal more protein flexibility and its importance in docking. In a validation study of the SLIDE docking tool, Zavodszky and Kuhn examined how many binding events could be modeled if an apo protein structure was only allowed minimal side-chain rotations.[166, 167] They compared their SLIDE docking tool to rigid docking with 20 different proteins (having 63 holo structures and 20 apo structures), where the backbone RMSD between the apo and holo structures $\leq$ 0.5 Å (thus no backbone changes would be necessary to dock the ligand). Only minimal side-chain changes were needed. SLIDE was able to dock all of the ligands within 2.5 Å RMSD of the crystal-structure pose while rigid docking only worked for 32 of the 63 structures. SLIDE changed 94% of the side chains by $< 45^\circ$ and 82% of the side chains less than 15°. This range of movement used in SLIDE can be compared with the natural variation observed among different holo crystal structures. Among the holo crystal structures in their set, 90% of the side chains changed by $< 45^\circ$, and 75% changed by $< 45^\circ$. Thus, small changes are typical, but more importantly, they are critical for accurate results in half of their studied protein structures.

Zhao, Goodsell, and Olson examined flexibility differences between amino acids. They examined the variation of χ_1 angles among different apo structures of the same protein to establish limits of natural variation in side-chain χ_1 of each amino acid. The authors established ranges for each amino acid that represent 90% of the observed conformations. Ile, Thr, Asn, Asp, and large aromatics showed limited flexibility, but Ser, Lys, Arg, Met, Gln and Glu were very flexible.

Najmanovich *et al* examined side-chain flexibility upon ligand binding with their BPK database of 221 proteins containing 523 holo structures matched with 255 apo structures.[168] Overall, 94.4% of all χ_1 angles changed less than 60°. In 40% of the

apo-holo protein pairs, none of the χ_1 values differed by more than 60°. However, the other 60% had at least one χ_1 undergo a large conformation change beyond 60°. Rotations of 60° or greater are significant enough that most rigid docking will fail, but more importantly, many movements that are less than 60° will still be problematic. Therefore, less than 40% of these structures can be adequately treated without including flexibility. This study then showed that no correlation could be found between backbone movements (measured in the largest C_α displacement) and side-chain flexibility (measured as the fraction of side chains undergoing a change of $\geq 60°$. This easily explains cases where C_α RMSD implies a protein is rigid, but χ-angle analysis reveals a flexible binding site.

These studies reveal that many residues in binding sites do not undergo significant rearrangement upon ligand binding, but there are often a few key residues with significant flexibility. However, some of the studies noted above are limited to very small sets of proteins. Sampling issues become important with very small datasets. Additionally, some of the studies contrast changes in the protein upon ligand binding (apo to holo) with changes seen between two structures of the protein either with or without ligand (holo-holo comparisons or apo-apo comparisons), but not both. While changes from ligand binding have been seen, they have not been appropriately separated from inherent variation observed in proteins.

This study aims to address the issue of protein flexibility upon ligand binding, employing a large dataset and focusing on contrasting inherent flexibility to changes upon binding. Because each protein in the dataset has at least two holo and two apo structures, it can compare the observed changes to variation seen among proteins with ligands (holo) as well as with variation seen among proteins without ligands (apo). It uses a large and carefully created dataset so that the observed differences can be statistically quantified. This study describes a comprehensive set of 214 proteins, represented by 1276 holo and 983 apo protein crystal structures. We describe statistically significant differences in flexibility upon ligand binding, looking for correlations with other properties such as ligand size, crystal-structure resolution, enzymatic function, and catalytic composition of residues.

Additional questions this study addresses are: first, are catalytic residues more or less rigid than noncatalytic residues? Second, is there a difference in the type of binding event, i.e., is there a difference in flexibility between enzymes (catalytic) and nonenzymes (noncatalytic binding pockets)? Third, does the size of the ligand influence flexibility?

4.2 Methods

4.2.1 Holo Dataset

Binding MOAD was used as a source of high quality protein-ligand complexes that have a maximum of 2.5 Å resolution.[169] Biologically relevant ligands were differentiated from opportunistic binders in the crystal structure (e.g. salts, buffers, phosphate ions). Furthermore, this dataset has ensured that none of the ligands were covalently attached. For biologically relevant ligands, the pocket was defined to include all protein residues that were within a 4 Å radius around the biologically relevant ligands, which should capture hydrogen-bonding and van der Waals interactions. Structures with more than one valid ligand were excluded from this dataset in favor of binary protein-ligand complexes to ensure that only one pocket was analyzed in each protein.

One unique feature about this set is the size of the ligands involved. This dataset allows for ligands that are composed of more than one HETATM group from the crystal structure. This study allowed peptides up to 10 amino acids, nucleotides up to 4 in length, as well as other multipart ligands.

Each structure in the holo set was clustered based on sequence identity using stringent criteria of 100% sequence identity to focus solely on the effect of ligand binding and not the influence of a mutation. Sequence identity between structures was determined using BLAST.[170]

4.2.2 Apo Dataset

A set of apo structures was identified by screening the PDB for structures of 2.5 Å resolution or better and then identifying structures without any HET groups (except for water) or only having HET groups that are not biologically relevant (like crystallographic additives). Acceptable additives were restricted to HET groups of 5 atoms or less, as well as a molecular weight of 100 Daltons or less. Each HET group was inspected for chemical appropriateness.

Apo structures were matched to holo structures by aligning sequences and requiring 100% sequence identity. Proteins that did not have at least two holo proteins and two apo proteins were excluded from the datasets.

4.2.3 Active-Site Identification

As most of the structures for a protein had different ligands bound and each had a different set of residues near the ligand, the intersection of all sets of residues was used to identify the binding pocket for the protein. A subset of residues identified as catalytically active was created using the Catalytic Site Atlas (CSA), version 2.2.9.[57] The CSA is a set of residues that have been annotated as having direct catalytic function in enzymes. It was built using hand annotation of sites from literature and expanded by using sequence analysis to identify residues in homologous structures[57].

4.2.4 RMSD Calculations

In order to compare the overall structural similarity of all the structures of a protein, we calculated an average RMSD. Since RMSDs are pairwise comparisons, this involved choosing a structure and comparing all other structures of that same protein to this structure. The structure that yielded the smallest average RMSD for the all of the protein structures was used (so that if any other structure was chosen, it would yield a higher average RMSD). Two different RMSD calculations were done: one using all backbone atoms of the protein and another using all atoms of residues in the active site. To examine the flexibility of the side chains, χ_1 was measured for residues with comparable torsion angles except for Gly, Ala, and Pro. The variation for a given residue was measured by determining the maximum range of χ_1 values observed.

4.2.5 Ligand Size

The molecular weight (MW) for each ligand was calculated according to the formula for the ligand in the PDB file. Some ligands are composed of more than a single HET component and were appropriately treated as one large molecule. For example, the inhibitor TER-117, in the PDB structure 10GS of Human Glutathione S-Transferase, is comprised of the HET groups 'GLU BCS PG9'.

4.2.6 Permutation Test Based on the Bootstrap Method

To assess if an observed difference between two groups is sufficient to reject the null hypothesis that the two groups have identical distributions, a permutation test based on bootstrap sampling was used. To perform a permutation test, first, a measurement was

taken of the difference between the two groups. This measurement is called the test statistic T_{orig}. Second, the observations were resampled into two new groups by pooling all of the observations and then randomly assigning them into two new groups with the same size of the two original groups. Third, the difference between the two new resampled groups was measured and recorded as T_{new}. The second and third steps are repeated a large number of times, T_{total}. For all tests in this study, 10,000 resamples were taken. A two-sided *p*-value was determined by taking the total number of times that the absolute value of T_{new} was greater than or equal to the absolute value of the original measurement T_{orig}, divided by the total number of resamples taken, T_{total}. This *p*-value estimates the likelihood of observing the measurement in a random sample, or in other words, the probability that the two groups are the same and the observed difference is a random anomaly.

4.3 Results and Discussion

4.3.1 Dataset Properties

The 2007 Binding MOAD release has 9836 protein-ligand complexes. When these 9836 structures are clustered at 100% sequence identity, there are 5526 different proteins. Upon filtering for proteins with at least 2 holo structures and 2 apo structures, this dataset reduces to 214 different proteins, represented by 1276 holo structures and 983 holo structures.

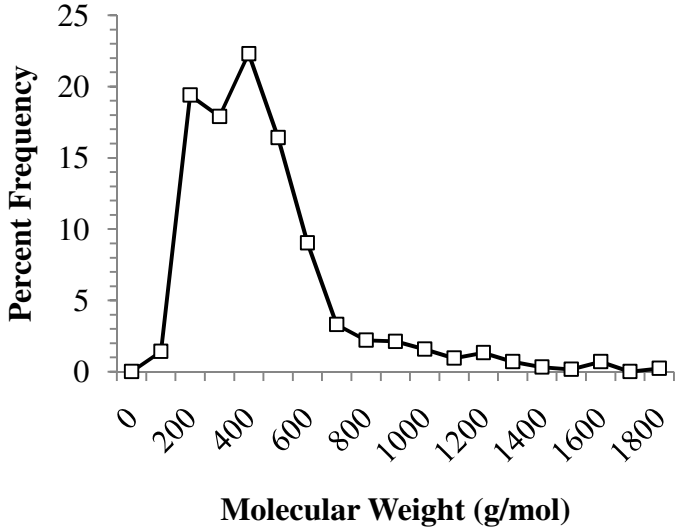
The protein with the most holo structures is tyrosine phosphatase 1B with 119 holo structures. The second largest is thrombin with 46 holo structures. The protein with the most apo structures is lysozyme with 175 apo structures, followed by ribonuclease A with 73 apo structures.

There is a range of ligand sizes (see Figure 4.1). Overall, 77% of the ligands are under 500 g/mol, and 92% of the ligands weigh less than 800 g/mol. The heaviest ligand is bivalent nitrophenol-galactoside ligand BV4 in crystal structure 1RF2, which weighs 1795 g/mol (and inhibits the protein at an IC_{50} of 17.0 μM). The smallest ligand is the carbonic anhydrase II inhibitor, 1,2,4-triazol, which weights 69 g/mol. The average MW of the ligands is 398 g/mol.

4.3.2 Resolution

Based on the observation that binding a ligand appears to rigidify a protein, which might in turn lead to better resolved crystal structures, correlation between resolution and binding

Figure 4.1 Distribution of Ligand Sizes



affinity was examined using all binding data available for holo proteins in Binding MOAD. No correlation was found between how tightly a ligand binds and the resolution of a structure (see Figure 4.2).

Resolution was investigated to see if it correlated with flexibility of apo and holo proteins. Figure 4.3 shows that apo proteins tend to have the same resolution as holo proteins (an average of 1.94 Å for holo and 1.94 Å for apo with no correlation between them). There are 117 structures that are better resolved as holo proteins than apo, and there are 97 proteins that are better resolved in their apo form. The slight difference suggests that there is a light influence, at best, for proteins when bound to a ligand.

4.3.3 RMSD

Average backbone RMSD calculations were measured for each protein, as shown in Table 4.2. By comparing the backbone RMSD among apo structures to the backbone RMSD for holo structures for the same protein, it is shown that the apo proteins are more varied structurally (See Figure 4.4). In 72 cases, the apo structures showed at least 0.2 Å higher backbone RMSD than holo structures (points above top-most dashed line). In only 22 cases, did the holo structures exhibit a backbone RMSD more than 0.2 Å greater than the apo (points to the right of the lower-most dashed line). In 120 of the 214 proteins, the holo and apo structures were within 0.2 Å RMSD (points in between the two dashed lines in Figure 4.4). Most of the backbone RMSD values are less than 1 Å, with a significant portion being

Figure 4.2 Resolution versus Free Energy of Binding

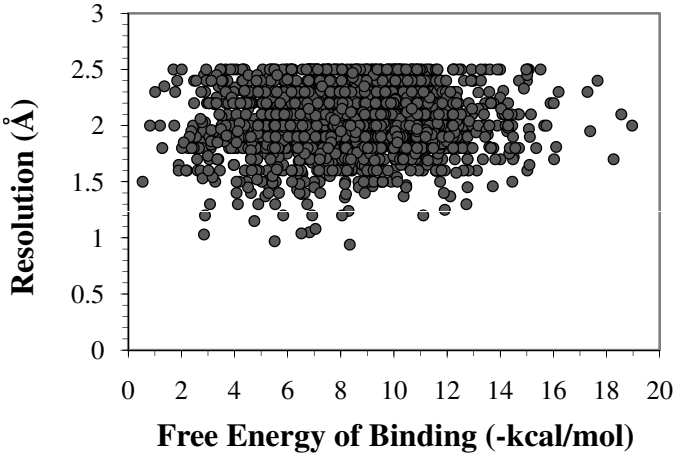
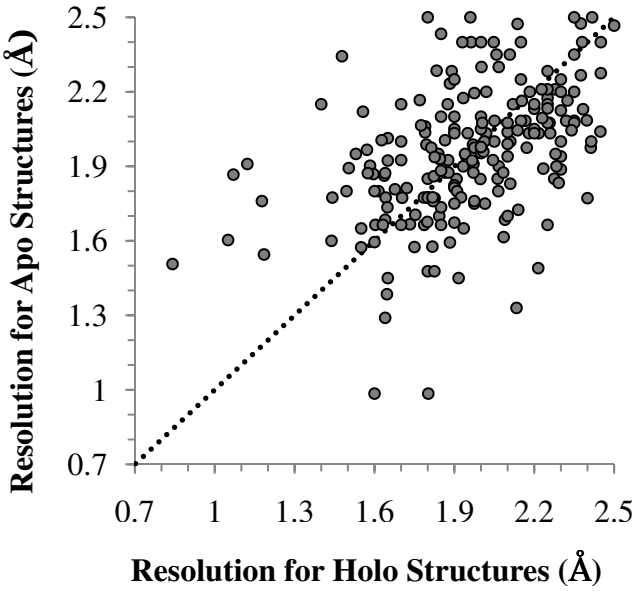


Figure 4.3 Holo Resolution versus Apo Resolution



less than 0.5 Å.

For a given protein, the backbone RMSD measured between apo and holo is generally smaller than the backbone RMSD across apo structures and larger than the backbone RMSD across holo structures (See Figure 4.5 and 4.6 as well as Table 4.2). This suggests that the holo structures are sampling a subset of the apo structures' conformational space, else the backbone RMSD for all structures would be greater than the backbone RMSD for apo structures. There are implications in choosing holo structures over apo structures in docking studies. As holo structures are more similar to each other than to the apo structure, a holo

structure should be better than an apo structure for docking.

For holo structures, the all-atom active-site RMSD measurements are smaller than backbone RMSD measurements (see Table 4.2). For apo structures, the all-atom active-site RMSD measurements are significantly smaller than the backbone RMSD measurements. The small all-atom active site RMSD means there not much structural variation in active site among holo structures or in active sites among apo structures. Figure 4.7 shows that 178 proteins have the holo structure all-atom active-site RMSD within 0.2 Å of the all-atom active-site RMSD for apo structures. However, there is a large difference in the active-site RMSD between holo and apo structures, as shown in Table 4.2 and Figures 4.8 and 4.9. The active-site RMSD is larger between apo and holo structures compared to the active site RMSD among apo or holo structures alone, suggesting that the holo and apo structures occupy different conformational space. The difference between apo and holo active sites supports the concept of ligand binding inducing a fit or inducing strain in the binding site.[19] While ligand binding does not generally induce significant changes to the backbone, as seen by C_α RMSD alignments, it has a significant impact on the side chains.

Figure 4.4 Average backbone RMSD measurements for proteins with ligands (holo) versus proteins without ligands (apo). Data shown with a 1.5 Å RMSD cutoff. There are five structures that have holo RMSDs greater than 1.5 Å and two holo structures have RMSDs greater than 1.5 Å. The 120 points that fall in between the dashed lines have very little difference in apo versus holo RMSD. Apo structures show more structural variation for 72 proteins, and only 22 show more variation across ligand bound structures.

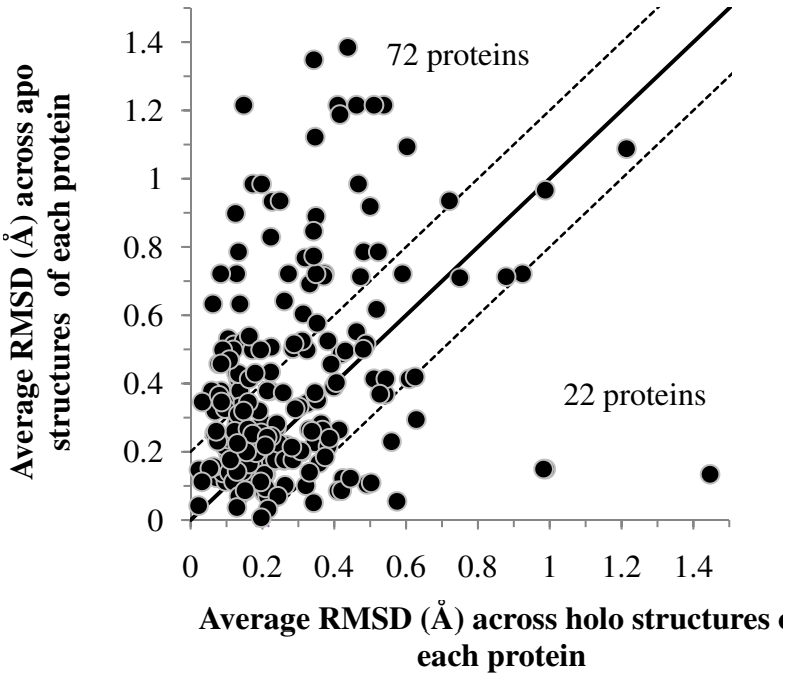
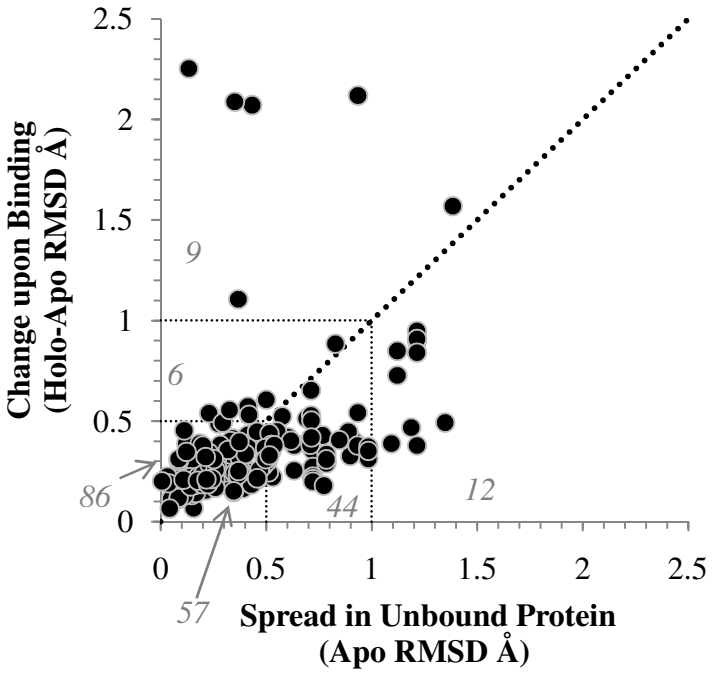


Table 4.2 Average RMSD Measurements

	backbone RMSD (Å)	active-site RMSD (Å)
Holo structures	0.34	0.19
Apo structures	0.46	0.25
Apo to Holo	0.42	0.60

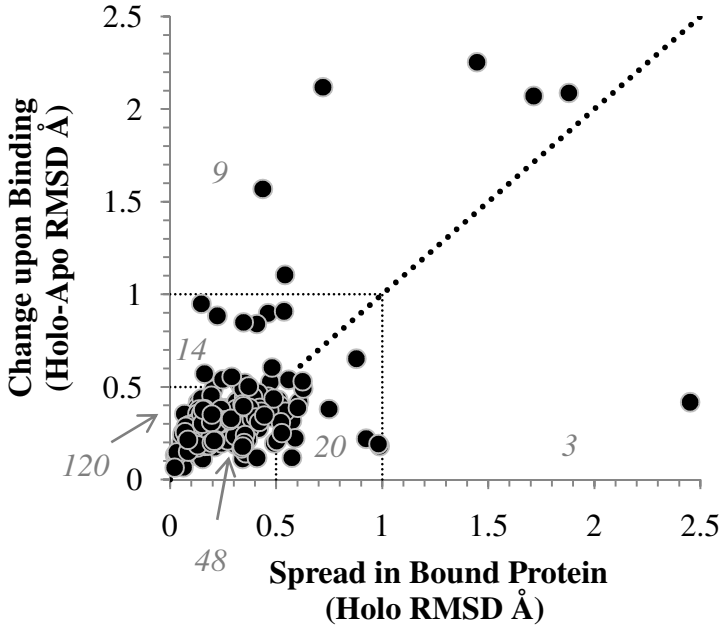
Figure 4.5 Changes observed upon binding versus spread across unbound structures, with cutoffs at 2.5 Å. The number of points in each section is labeled in gray. For proteins where the RMSD of apo structures is larger than the RMSD between apo and holo, there are 57, 44, and 12 proteins with apo RMSDs under 0.5 Å, between 0.5 Å and 1.0 Å, and above 1.0 Å, respectively. For proteins where the RMSD of apo structures is smaller than the RMSD between apo and holo, there are 86, 6, and 9 proteins with apo RMSDs under 0.5 Å, between 0.5 Å and 1.0 Å, and above 1.0 Å, respectively.



4.3.4 Comparing Holo and Apo Structures

To focus solely on side-chain behaviour, we examined the range of χ_1 angles for residues near the ligand pocket (see Figure 4.10). The ranges of χ_1 were compared between apo and holo structures (see Figure 4.11). The apo structures exhibited a wider range of dihedral angles than holo (see figure 4.11). This is supported by a study of B-factors between holo and apo structures where 71% become less mobile upon ligand binding, and 29% become

Figure 4.6 Changes observed upon binding versus spread across bound structures with ligand with cutoffs at 2.5 Å. The number of points in each section is labeled in gray. For proteins where the RMSD of holo structures is larger than the RMSD between apo and holo, there are 48, 20, and 3 proteins with apo RMSDs under 0.5 Å, between 0.5 Å and 1.0 Å, and above 1.0 Å, respectively. For proteins where the RMSD of holo structures is smaller than the RMSD between apo and holo, there are 120, 14, as 9 structures with holo RMSDs under 0.5 Å, between 0.5 Å and 1.0 Å, and above 1.0 Å, respectively.



more mobile.[171]

The difference in flexibility between holo and apo structures is supported by statistical tests. The two-tailed permutation test estimates the probability that the difference in flexibility between holo and apo structures is random at p=0.0487 when 60° threshold is used, and p=0.0803 when a 30° threshold is used. The fact that the 30° threshold has a smaller p-value is reasonable because using the stricter 30° threshold will categorize more residues as being flexible compared to a wider 60° threshold. Thus, with more residues being categorized as "more flexible" the 30° threshold will show a slightly smaller difference between the two populations.

The average residue in holo proteins exhibits a χ_1 range of 27.2°, while the χ_1 range in apo proteins averages 34.0° (see Figure 4.12). The average range when combining all structures (including both holo and apo) is 54.9°. The larger range of χ_1 values for all structures (compared to solely apo or holo structures, see Figures 4.13, 4.14) further supports the view that ligand binding is inducing strain on the side chains, such that holo structures are sampling additional orientations no sampled in the apo structures.

Figure 4.7 All-atom RMSD of active sites in holo and apo structures. The points that fall in between the gray, dashed lines have very little difference in apo versus holo active-site RMSD. There are 23 proteins that show more structural variation without ligand compared to bound. There are 13 proteins that show more structural variation when bound to ligand than ligand-free.

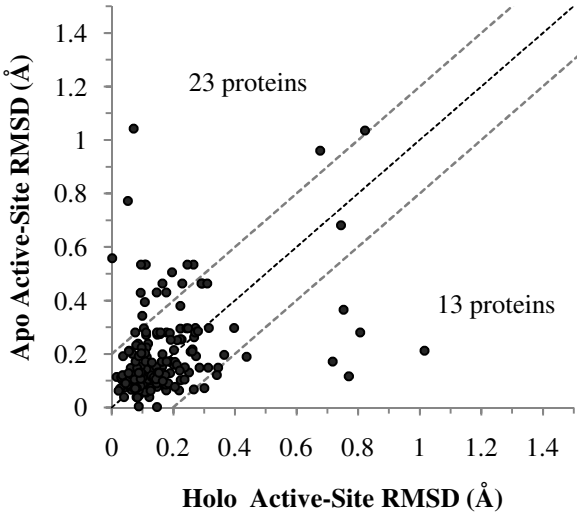
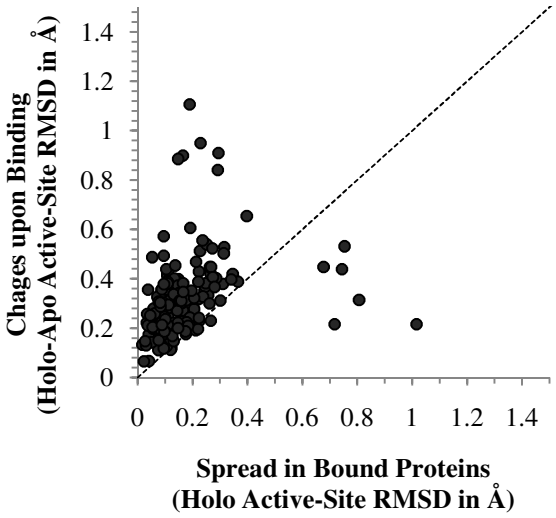


Figure 4.8 Changes observed upon binding versus spread in bound structures (all-atom, active-site RMSD)



4.3.5 Influence of Number of Structures Representing a Protein and Ligand Size

It is possible that the larger χ_1 ranges are seen when the apo and holo structures are combined. The influence of how many structures represent a protein was analyzed (see Figures

Figure 4.9 Changes observed upon binding versus spread in apo structures (all-atom, active-site RMSD)

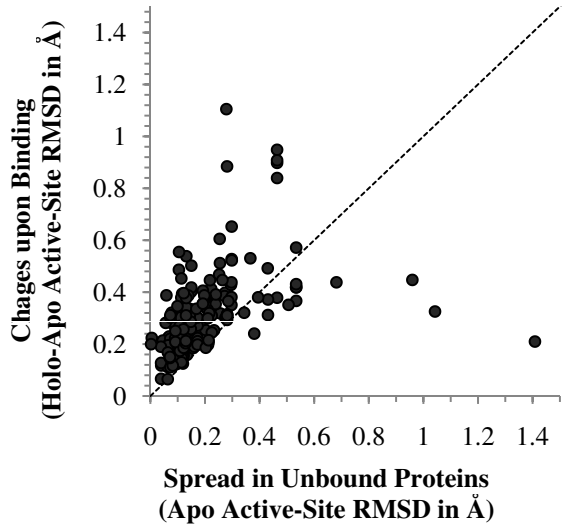
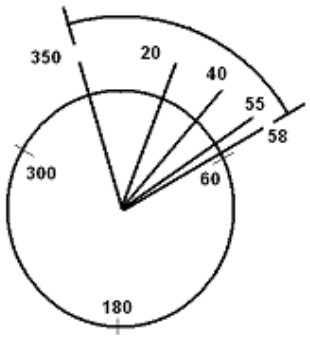


Figure 4.10 Measuring the χ_1 angle range. A Fischer-like projection, illustrating the variation in dihedral angle for a given residue. Five different crystal structures of the same protein may have five different dihedral angles for a given residue. Here, five different dihedral angles are represented with a range of 68°.



4.15, 4.16). For both apo and holo, proteins with more structures showed more flexibility than proteins represented with fewer structures. There are two possible causes. First, the proteins with more structures simply have more data to analyze and are more likely to reveal variation at a given residue (thus a protein with 20 structures will have 20 side-chain conformations of a given residue for analysis, but a protein with only two structures has only two conformations per residue). Second, this study defines active-site residues as the union of all residues within 4 Å of a ligand bound in any of its holo structures. In some cases, proteins with more structures may have more residues defined as part of the binding site. This allows for averaging conformational behavior over more χ angles. It is unlikely that the binding-site definition is problematic because the site is defined the same in the apo

Figure 4.11 The range of χ_1 angles for binding-site residues is compared between apo structures (gray circles) and ligand-bound, holo structures (black diamonds). Vertical lines emphasize that 70.0% of the residues in apo structures and 75.1% in holo structures had χ_1 ranges $\leq 10°$, while 91.4% of the residues in apo structures and 93.0% in holo structures had χ_1 ranges $\leq 60°$.

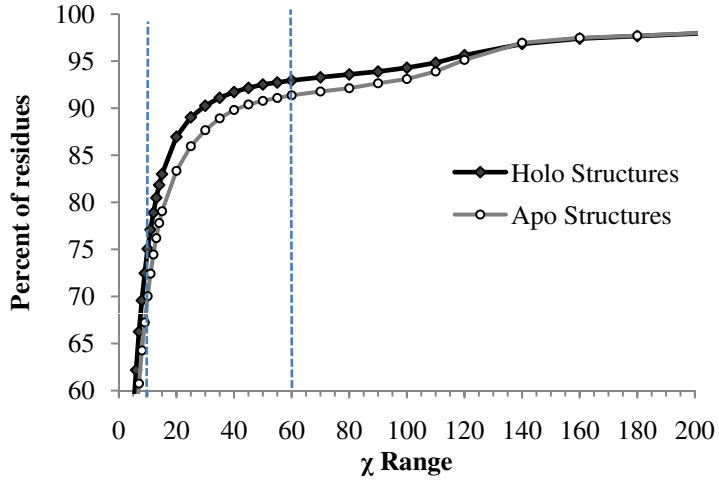
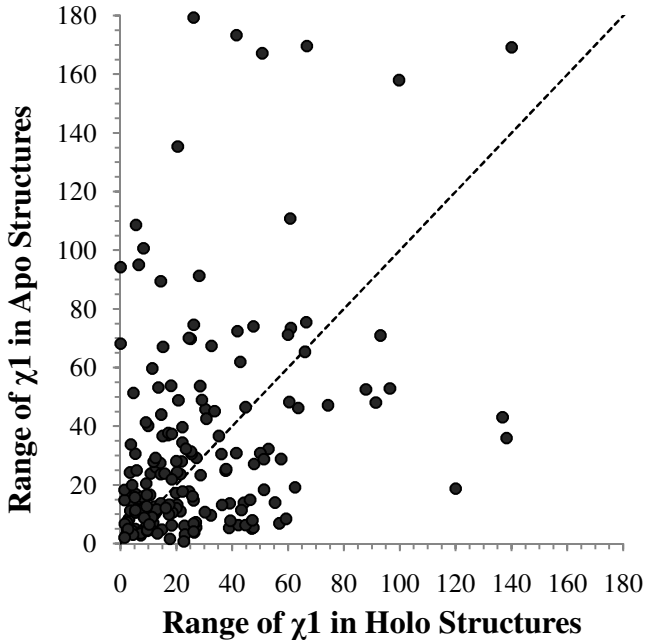


Figure 4.12 Average range of active-site χ_1 values in holo versus apo structures for each protein.



and holo structures of the same protein. Furthermore, if we changed the definition of the binding pocket such that it required that at least two structures to have the same residue with within 4 Å of a ligand, it did not significantly change the results. When greater flexibility is observed in sets with more available structures, does this simply reflect more chances to

Figure 4.13 Average range of χ_1 values in apo structures vs all structures for each protein.

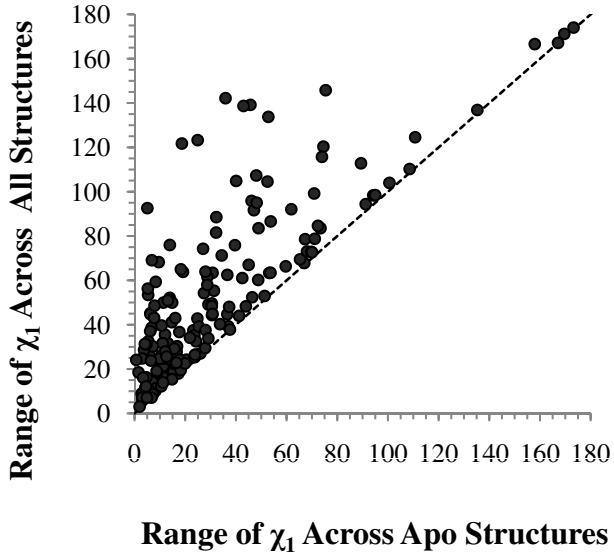
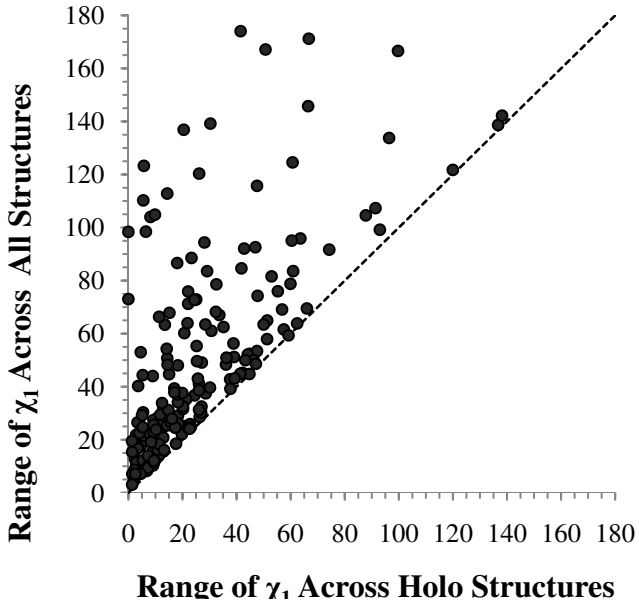


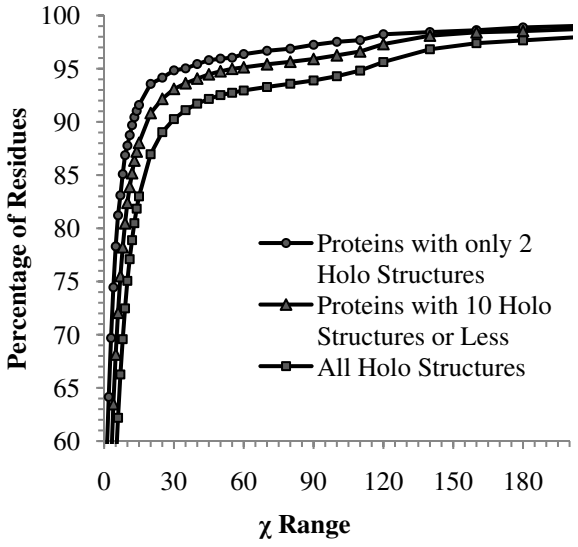
Figure 4.14 Range of χ_1 values in holo structures versus all structures.



observe rotameric changes, or do sets with many structures happen only when a protein is flexible and more unique structures are solved and published? While the answer is unclear, it does point to the need for large datasets, rather than single apo-holo pairs, for accurate insights into protein flexibility.

The effect of ligand size on flexibility was examined. However, no properties had any significant correlation with the average ligand size. For example, there was no correlation between ligand size and percentage of residues with movements above any threshold, between average ligand size and average RMSD, or between average ligand size and average resolution. (Data not shown)

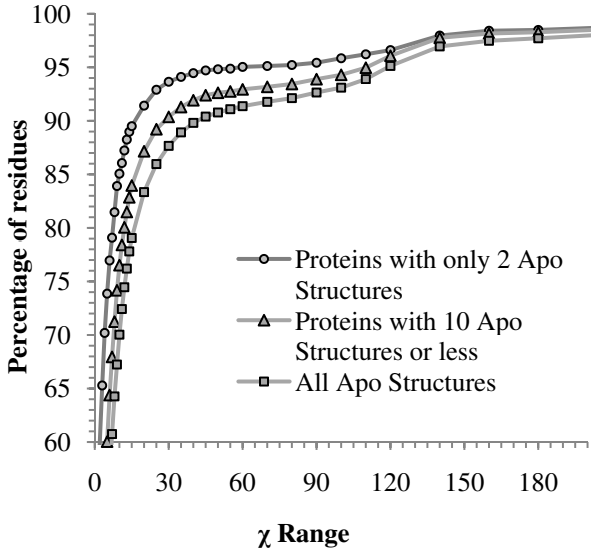
Figure 4.15 Range of χ_1 angles for binding-site residues of proteins that have only 2 holo structures, for proteins that have 10 holo structures or less, and for all proteins (regardless of the number of holo structures).



4.3.6 Influence of Amino-Acid Composition

Figures 4.17, 4.18, and 4.19 shows the inherent flexibility of amino acids (e.g., 81% percent of Arg residues are found within 30° of each other). They show the propensity that each amino acid has for rotameric variation. Overall, the trend follows (from least flexbile to most flexible) Trp, Phe, Leu, His, Ile, Tyr, Thr $<$ Asp, Lys, Gln, Asn $<$ Met, Val, Cys, Glu, Arg. Other studies have shown very similar trends, although the ranking is not exact.[172, 168] The trend does match the pattern of large hydrophobic residues being more stericly constrained and large polar or charged residues being more flexible. It is interesting that Asp, Lys, and Gln were not shown to be more flexible. Perhaps there is a connection with Asp and Lys being common catalytic residues.

Figure 4.16 Range of χ_1 angles for binding-site residues of proteins that have only 2 apo structures, for proteins that have 10 apo structures or less, and for all proteins (regardless of the number of apo structures).



4.3.7 Influence of Catalytic Residues

Flexibility of catalytic residues was measured and compared to the flexibility of noncatalytic residues. Catalytic residues in holo proteins were not statistically different from noncatalytic residues in holo proteins (p=0.7758 using 60°). The trend holds likewise in apo proteins, p=0.9546 that catlytic residues are the same as noncatlytic residues. The trend holds using a 30° threshold.

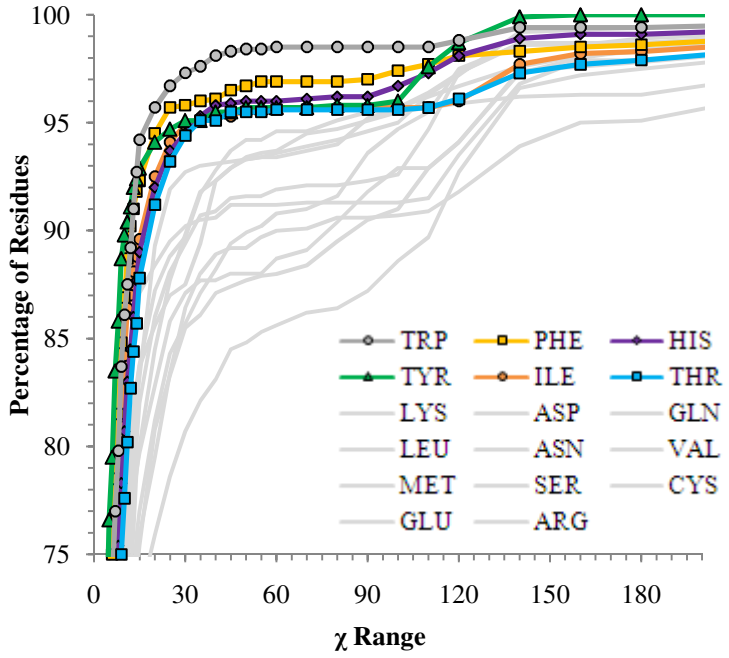
Gutteridge and Thornton also found no difference in flexibility between catalytic and noncatalytic residues, but did find that noncatalytic residues tended to undergo more backbone motions.[164]

4.3.8 Influence of Protein Function

Enzymes and nonenzymes are functionally different. Our previous work showed that nonenzymes have greater ligand efficiencies than enzymes, where ligand efficiency is defined as the binding affinity divided by the number of heavy atoms in the ligand.[173] We hypothesized that this difference is due to different evolutionary pressures: nonenzymes need to be sensitive to low concentrations of signaling molecules whereas enzymes need to bind molecules, change them, and then release them.

To see if there is a difference in their flexibility, the proteins were divided into enzymes

Figure 4.17 Percentage of residues within a χ_1 range.



and nonenzymes based on their Enzyme Classification and annotated function. There were 160 enzymes and 54 nonenzymes. No correlation with flexibility and function was found. While highly efficient enzyme-ligand complexes have a higher propensity to have certain residues in the active site compared to highly efficient nonenzyme ligand complexes, there was not a clear difference in the flexibility between the two subsets of amino acids. Highly efficient enzymes have Thr, Asp, and Val in the binding site and nonenzyme-ligand complexes in turn display a higher propensity to have Tyr, Trp, Leu.[173]

Enzymes followed the same pattern as of proteins as a whole, for backbone RMSD, active-site RMSD, and χ_1 range patterns (see Table 4.3). This is very reasonable, considering that most of the proteins in this study are enzymes. Nonenzymes follow the pattern for active-site RMSD and the χ_1 range. Interestingly, nonenzymes had a larger backbone RMSD between apo and holo structures than among apo structures. Nonenzymes also had larger backbone RMSDs, active-site RMSDs, and χ_1 ranges than enzymes, for holo, apo, and holo to apo comparisons.

Figure 4.18 Percentage of residues within a χ_1 range.

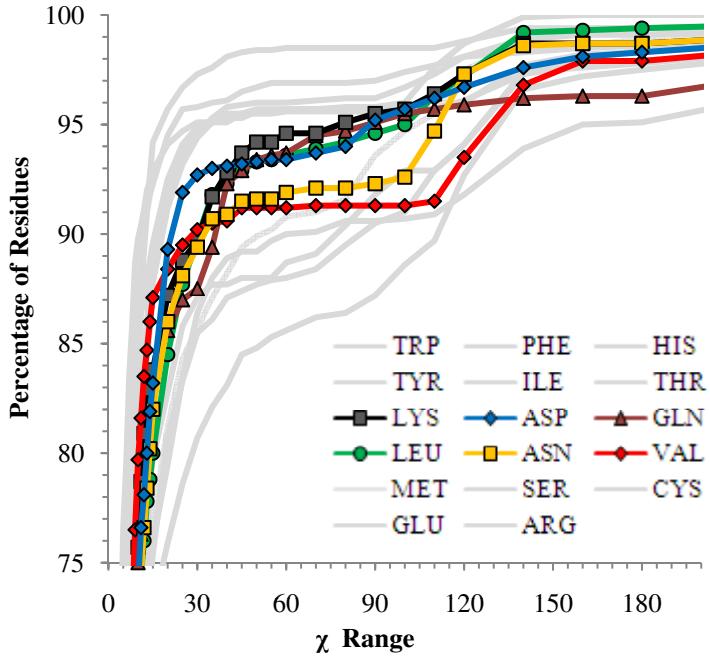


Table 4.3 Variation seen among and between holo and apo structures for both enzymes and nonenzymes.

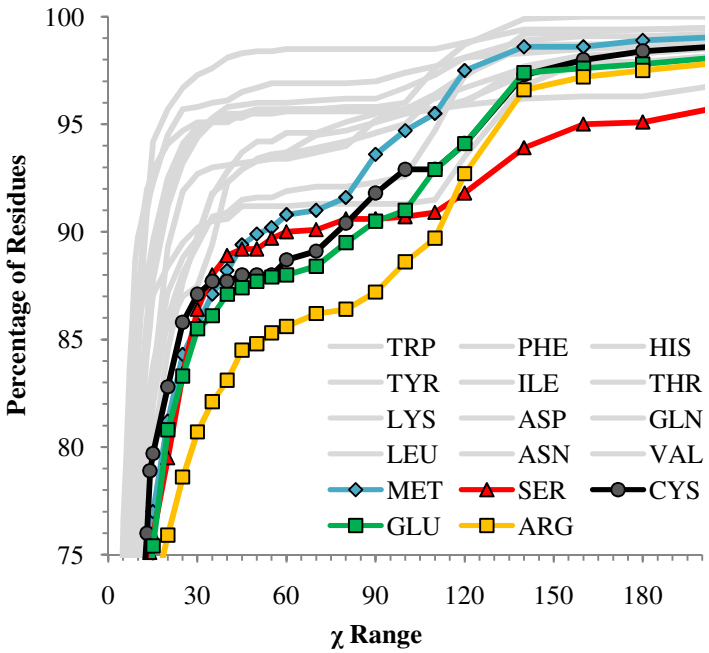
		backbone RMSD (Å)	active-site RMSD (Å)	χ_1 range
Enzymes	holo	0.27	0.17	26.2
	apo	0.39	0.18	31.6
	holo-apo	0.32	0.46	50.1
Nonenzymes	holo	0.53	0.24	32.4
	apo	0.62	0.38	38.8
	holo-apo	0.74	1.03	64.8

4.4 Conclusion

Understanding protein flexibility is important in drug design, especially when crystal structures are widely used as models for binding prediction[174]. This study examines how ligand binding influences protein flexibility. More specifically, it uses a large collection of proteins that have at least two holo and two apo structures, to examine what backbone and active-site differences are observed in inherent variation among holo or apo structures and what differences come from ligand binding.

It was been shown that ligand-free structures have a small degree of natural backbone variation, as measured by backbone RMSD, and that most holo structures exhibit smaller

Figure 4.19 Percentage of residues within a χ_1 range.



backbone RMSD values. The natural apo backbone variation becomes constricted upon ligand binding, resulting in more backbone structural similarity among holo structures. Thus, any given holo structure may be a better starting point for modeling ligand binding than most apo structures. This is especially true considering that the of the size of the ligands did not appear to influence flexibility.

Freire and Luque have illustrated how energy from ligand binding is not necessarily uniformly distributed throughout the protein; binding sites can be regions of both high and low stability.[14] We examined if there was a difference in flexibility between catalytic and catalytic residues in active sites, as well as between enzymes (catalytic) and nonenzymes (noncatalytic binding pockets). No differences, however, could be found based on these classifications.

Shifting from examination of the global backbone changes to local influence of binding in the active site revealed another level of ligand binding induced effects. Apo structure are observed to have a certain range of flexibility in their active sites, just as holo structures have a similar, but smaller degree of variation among their active sites. However, there is greater variation between these two groups than there is within either group by themselves. This is evidence of an induced fit, where ligand binding induces active-site side chains to occupy a different conformational space than before. This conformational change upon ligand binding is supported by both active-site all-atom RMSD and chi range analysis.

Ligand binding has a demonstrable effect on protein flexibility, at several levels. The influence on protein backbones becomes apparent, for example, upon looking at natural flexibility across a broad selection of proteins. The insights in this study have come from examining a panoply of holo and apo structures - highlighting the important contributions that large datasets, such as Binding MOAD, can make in science.

Chapter 5

A Novel Test Set for Evaluating Scoring Functions

5.1 Introduction

Protein-ligand structure databases are important tools for studying principles of molecular recognition. Binding MOAD, for example, has helped yield new insight into protein-ligand binding, ranging from fundamental differences between enzymes and nonenzymes to ligand-induced changes in flexibility.[173] Binding MOAD is valuable not only for generating new knowledge, but also as a yardstick for testing current understanding and hypotheses. It can serve as gold standard collection of known protein-ligand structures for evaluating and improving algorithms used in molecular docking. Here, we demonstrate its use in creating a new test set to evaluate scoring functions, one that poses a new question, "can scoring functions distinguish biologically relevant binding across diverse proteins?"

Molecular docking has become an increasingly important computational tool in modern structure-based drug design (see refs [175, 152, 153, 176, 177, 178] for review). Given the three-dimensional structure of a protein, the molecular docking process starts with sampling possible ligand orientations and conformations (referred to as modes) at the selected site of the protein target and then ranks these modes according to their scores calculated with a scoring function. The development of accurate scoring functions to evaluate putative modes is a critical and challenging element in molecular docking. For years, different scoring functions have been developed that boast different computational speeds and accuracy (see Table 5.1). Roughly, these scoring functions can be grouped into three categories according to their derivation: force-field based, empirical, and knowledge based.

Force-field based scoring functions are fully or partially evaluated on a set of force-field parameters derived from both experimental work and quantum mechanical calculations to

describe the interactions among atoms.[179, 24] When considering the multitude of ways that explicit water molecules can compliment binding (see ref [180] and references therein), conformational sampling and force-field based scoring functions are computationally too expensive to be used in virtual database screening. As an alternative, the solvent effect can be implicitly considered by the Poisson-Boltzmann model (e.g., [181, 182, 183, 184]) or generalized-Born model (see [185, 186] for review) in post-docking scoring (e.g., [187, 188, 189]). The most simplified method for modeling the solvent effect is to use a distance-dependent dielectric constant to calculate the electrostatic interaction energy term [26], which can be directly used to speed the docking process at the expense of accuracy.

A second category is empirical scoring functions whose parameters are derived by reproducing the binding affinities of a training set of protein-ligand complexes with known three-dimensional structures (e.g., [190, 191, 35, 32, 33, 192, 193]). Compared to force-field based scoring functions, empirical scoring functions score protein-ligand complexes quicker because of their relatively simple energy terms. The generality of an empirical scoring function is typically restricted by the composition of its training set.

The third kind of scoring functions are the knowledge-based scoring functions [194, 195, 196, 197], in which an inverse Boltzmann relationship is used to determine pairwise energy potentials, directly converted from the occurrence frequencies between atom pairs in a database of protein-ligand structures.[198, 41, 39, 199, 200, 201, 202, 203, 42, 38] The derived pair potentials try to embody all the effects that govern ligand binding such as electrostatic interactions, van der Waals interactions, hydrophobic effect, desolvation penalties, etc. Knowledge-based scoring functions have a good balance between accuracy and speed. Compared to empirical scoring functions, knowledge-base scoring functions can be more general as a result of larger and more diverse training sets of protein-ligand structures available from the Protein Data Bank (PDB) because any structure can be used even if binding affinity data is unknown.[61] The pair-potential feature of the knowledge-based scoring functions also makes the scoring process as fast as the empirical scoring functions.

Currently, there are three common criteria that are used to evaluate a scoring function.[38] The first criterion is binding-mode prediction, how closely a predicted ligand-binding mode resembles the experimental structure. The second criterion is binding-affinity prediction, whether or not the scoring function can rank order compounds by affinity or reproduce the experimentally determined binding data. The third criterion is enrichment in virtual database screening, whether or not the true inhibitors/binders can be ranked at the top of a large database of ligands according to their binding scores for a protein target. Most current scoring functions perform satisfactorily in one or two criteria [71]; however, it is challenging for a scoring function to perform well in all three.[38]

One common feature for the three above criteria is that they are designed to evaluate a scoring function on a single protein-ligand complex or a specific protein target without considering the biological types of the bound ligand. With the rapid development of proteomics projects, more and more protein-ligand structures are being determined experimentally and deposited in the PDB [61]. It is noticeable that many bound ligands in the PDB are biologically irrelevant; typical examples include additive molecules such as detergents for crystallization purposes or buffer molecules. The presence of these molecules bound to protein surfaces usually results from their high concentrations rather than from tight binding interactions (a case referred to as "opportunistic binders" or "invalid ligands").[169] Whether a scoring function is able to discern invalid ligands from weakly-bound, biologically relevant ligands is a new criterion proposed in the present work. It is desirable to extend scoring functions to evaluate protein binding sites for the determination of function or druggability of a pocket. This goal requires scoring functions to be able to discern biologically relevant binding events from opportunistic ones over a wide range of proteins. This can be particularly challenging if the biologically relevant binding is weak. An important counter issue is "appropriate failures" when additives in a true site should score well if they are chemically similar to the biologically relevant ligand.

In the present work, a diverse benchmark of valid and invalid protein-ligand complexes from the PDB is presented. Four different scoring functions, representing different categories, are used to test this new benchmark. The influence of including entropic penalties for rotatable bonds in ligands was also examined.

Table 5.1 Scoring functions and the size of training sets

Scoring Functions	Year	Complexes
Score1	1994	54
F-Score	1996	19
VALIDATE	1996	65
ChemScore	1997	112
ProteusScore	1997	82
Score2	1998	94
PMF	1999	225
BLEEP	1999	90
DrugScore	2000	83
SMoG	2002	119
HINT	2002	53
X-Score	2002	230
ITScore	2006	851

5.2 Materials and Methods

5.2.1 Scoring functions

We selected four scoring functions from different categories to test our new benchmark. They include an empirical scoring function, X-Score [37], a force field-based scoring function in DOCK 4.0 [26, 29], a semi-empirical force-field based scoring function in AutoDock 4.0 [28, 204], and a knowledge-based scoring function, ITScore in MDock [42, 38, 205, 206].

X-Score

The empirical scoring function X-Score includes three individual scoring functions of HSScore, HPScore, and HMScore as [37]

$$\begin{aligned}
\text{HSScore} &= \text{C}_{\text{VDW},1} \cdot \text{VDW} + \text{C}_{\text{H}-\text{bond},1} \cdot \text{HB} + \text{C}_{\text{hydrophobic},1} \cdot \text{HS} + \text{C}_{\text{rotor},1} \cdot \text{N}_{\text{tor}} + \text{C}_{0,1} \\
\text{HPScore} &= \text{C}_{\text{VDW},2} \cdot \text{VDW} + \text{C}_{\text{H}-\text{bond},2} \cdot \text{HB} + \text{C}_{\text{hydrophobic},2} \cdot \text{HP} + \text{C}_{\text{rotor},2} \cdot \text{N}_{\text{tor}} + \text{C}_{0,2} \\
\text{HMScore} &= \text{C}_{\text{VDW},3} \cdot \text{VDW} + \text{C}_{\text{H}-\text{bond},3} \cdot \text{HB} + \text{C}_{\text{hydrophobic},3} \cdot \text{HM} + \text{C}_{\text{rotor},3} \cdot \text{N}_{\text{tor}} + \text{C}_{0,3}
\end{aligned} \tag{5.1}$$

where the van der Waals (VDW) energy term is calculated by a Lennard-Jones 8-4 potential, the hydrogen-bonding term (HB) is obtained from the hydrogen bonds between protein and ligand, and the rotor term (RT) stands for the number of effective rotatable bonds in the ligand molecule. The HS, HP, and HM terms calculate the buried, hydrophobic molecular surface of the ligand, the pairwise hydrophobic atom-contact potential, and the microscopic match of hydrophobic ligand atoms to the binding pocket, respectively. The coefficients in the scoring functions were obtained by fitting the binding affinities of 200 protein-ligand complexes with known structures [37]. In the present study, we used the average of the scores from the three scoring functions in Eq. (5.1) to represent the X-Score of a protein-ligand complex.

AutoDock

The scoring function in AutoDock 4.0 is a semiempirical, force-field based scoring function which includes five energy terms [28, 204]

$$\begin{aligned}\Delta G &= W_{\text{vdw}} \cdot \sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) \\ &+ W_{\text{elec}} \cdot \sum_{i,j}\left(\frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}}\right) \\ &+ W_{\text{hbond}} \cdot \sum_{i,j} E(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) \\ &+ W_{\text{tor}} \cdot N_{\text{tor}} \\ &+ W_{\text{sol}} \cdot \sum_{i,j}(S_i V_j + S_j V_i) e^{(-r_{ij}^2/2\sigma^2)}\end{aligned} \tag{5.2}$$

where the first two energy terms are classic VDW and electrostatic interactions and have the same forms as the force-field scoring function in DOCK 4.0 [26, 29]. The third term stands for the contribution from hydrogen bonds between protein and ligand. The fourth term considers the loss of torsional entropy of a ligand upon binding in which N_{tor} is the number of rotatable bonds in the molecule. The last term describes the solvation effect. The weighting coefficients for the five energy terms were obtained by fitting the known binding constants of 188 protein-ligand complexes. [204]

DOCK

The scoring function in DOCK 4.0 [29] represents a typical force-field based scoring function whose energy parameters are taken from the Amber force field [26]. This scoring function includes VDW and electrostatic interaction energy components

$$E = \sum_i \sum_j \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}}\right) \tag{5.3}$$

where r_{ij} stands for the distance of protein atom i and ligand atom j, A_{ij} and B_{ij} are the VDW parameters, and q_i and q_j are the atomic charges. The effect of solvents is implicitly considered by using a distance-dependent dielectric constant $\varepsilon(r_{ij})$.

ITScore

ITScore is an iterative knowledge-based scoring function developed using a training database of 781 protein-ligand complexes structures from the PDB [42, 38], representing a set of effective pair potentials resulting from the overall effects of all binding factors. The binding score is calculated by summing up all the atomic pairs between protein atom i and ligand atom j as

$$E_{\text{ITScore}} = \sum_{i,j} u_{ij}(r) \tag{5.4}$$

where r is the distance between the atom pair ij. The effective pair potentials $u_{ij}(r)$ are iteratively derived until they can discriminate the native structures from decoys for 99% of the protein-ligand complexes in the training set. The ITScore scoring function has been implemented in MDock, a program for docking against an ensemble of protein structures. [205, 206]

Adding Torsional Entropic Penalties

X-Score and AutoDock contain terms that penalize a score for each rotatable bond in a ligand on the bases that restricting each torsion carries an entropic cost. Similar terms are not included in DOCK or ITScore. The additive nature of both DOCK and ITScore inherently bias large ligands to score well (a well-known limitation of *many* scoring routines). This caveat can be particularly problematic in our study because several additives in the decoy set are large detergents.

To investigate the effect of incorporating torsional entropy penalties for the ligands we calculated two set of binding scores with and without a torsional ligand term. To remove the torsional term from X-Score and AutoDock, we simply set the coefficients C_{rotor} and W_{tor} to zero in Eqs. (5.1) and (5.2). These are referred to as X-Score-tor and AutoDock-tor, respectively. To add a torsional term to DOCK and ITScore in a straight forward way, we added $w_{\text{tor}} \cdot N_{\text{tor}}$ to Eqs. (5.3) and (5.4) where N_{tor} is calculated by X-Score (its rotatable bound count) and w_{tor} (a scaling factor for the torsional penalty term) was simply set to 1 for this work. These are referred to as DOCK+tor and ITScore+tor. This was chosen for simplicity. Furthermore, we did not wish to unfairly bias the performance of DOCK and ITScore by explicitly fitting new parameters for this purpose.

5.2.2 Hit and Decoy Dataset

The test set was composed of protein-ligand crystal structures, organized into two groups: a hit set of 33 complexes (Table 5.2) and a decoy set of 30 complexes (Table 5.3). Hits have biologically relevant ligands. The decoys are extraneous molecules found in the crystal structures, such as buffers, detergents, and solvents. While the decoys are true binding events appropriately resolved in the crystal structure, they are weak, opportunistic binding events induced by the high concentrations in the crystalline environment. There are 59 different protein families in the dataset of 63 complexes, where a family is defined as a set of proteins that are at least 50% sequence identity to each other. The four families with more than one structure in the test set are chalcone synthase (1CGZ is a hit while 1D6F is a decoy), pim1 kinase (1YXX is a hit, while 1YXS is a decoy), protein kinase A (1APM and 1Q61 are both decoys), glycine amidinotransferase (9JDW and 7JDW are both hits). Affinity data is available for 15 of the 33 hits(45%). A broad range of affinities was desired, given the task at hand.

Table 5.2 Hits

PDB ID	Ligand	Binding Data	PDB ID	Ligand	Binding Data
1B0O	PLM		1T0O	GAL	
1CGZ	STL		1TR7	MPD	K_d=0.15μM
1CHM	CMS	K_i=0.22mM	1UTJ	ABN	K_d=0.144mM
1E02	UNA	IC_{50}=0.7μM	1VG0	GER	K_d=0.8nM
1EXF	GLY		1VYG	ACD	K_d=10nM
1FDQ	HXA	K_d=53.4nM	1W3J	OXZ	K_d=484nM
1FEN	AZE		1YFS	ALA	
1FJ4	TLM	IC_{50}=25μM	1YP0	PEF	
1HDC	CBO	K_i=1.0μM	1YXX	LI7	
1OTH	PAO	IC_{50}=100nM	2ACO	VCA	K_d=2.5μM
1P4F	DRG		2B77	3CA	
1PB9	4AX	K_i=240μM	2CCS	4BH	IC_{50}=8.2μM
1PEA	ACM		2J1S	FUL	
1POT	SPD	K_d=3.2μM	3YAS	ACN	
1PX8	XYP		7JDW	DAV	
1R9E	PGO		9JDW	ABA	
1RQ5	CTT				

Structures in the hit set were chosen from Binding MOAD, a database of high-quality protein-ligand crystal structures[115]. Each of these structures has a resolution better than 2.5 Å and a non-covalently attached ligand. All ligands in all the structures contained in

Table 5.3 Decoys. Those in real binding sites are noted with asterices as "acceptable failures"

PDB ID	Ligand	PDB ID	Ligand	PDB ID	Ligand	PDB ID	Ligand
1APM	OCT	1N2F*	DTT	1S2U*	PEG	2CJP*	PG4
1D5R*	TLA	1OLL	EDO	1SHV	MA4	2G47	DIO
1D6F*	B3P	1PK3	BME	1TTO	TRS	2GW5	IPA
1D7H*	DMS	1PPA*	ANL	1UN8	MYY	2I3A	BTB
1FZV	MPD	1Q44*	MLA	1YXS	IMD	2J8X	URE
1IZ2	SUM	1Q61	MG8	1ZR3*	MES	8CHO	P4C
1LI0	BCT	1QST*	EPE	229L*	GAI		
1LIH	PHN	1RJM	EP1	2BF3	HTO		

Binding MOAD are classified as valid (biologically relevant) or invalid (molecules such as salts, buffers, detergents, and solvent). Structures where cofactors were not interacting with a ligand were desired. Furthermore, all structures in both sets were restricted to a pH range of 6-8. This gave a limited range of biologically relevant pH where the protein setup could be more accurately automated. To create a suitable decoy set, the same requirements of resolution and non-covalently attached ligands was also used. If possible, decoys were chosen from PDB files that had only one small molecule in the structure. Multi-part ligands, peptides, and nucleotides were not considered as decoys. Ligands in the decoy set were chosen to be chemically similar to at least one ligand in the hit set. This resulted in both the hit and decoy sets having ligands with very similar distributions of size, SlogP, and logS. Figures 5.1 - 5.3 show that the ligands in both sets primarily ranged from 100-500 molecular weight, -4 to 8 SlogP, and -6 to 2 logS. (SlogP and logS were calculated using MOE.[127]) Finally, it was important to avoid comparing sets of well-buried hits to surface-bound additives. Figure 5.4 shows that the distributions of buried surface area (BSA) are very similar for the hits and decoys. This is critical because BSA is proportional to the number of protein-ligand contacts that will dictate a docked score. Every effort was made to choose complexes to obtain similar distributions in the degree of exposure for the small molecules in the hit and decoy sets. This proved to be difficult to accomplish, especially while simultaneously restricting the chemical characteristics to remain the same between the two sets. Figure 5.5 shows that the decoys have a distribution shifted $\sim$10% more exposed. There is a strong bias in the hits to be 0 - 25% exposed, but 6 complexes have hits that range from 25 - 60% exposed. Of the decoys, 10 complexes are in that range.

Figure 5.1 Histogram of Hit and Decoy Sizes

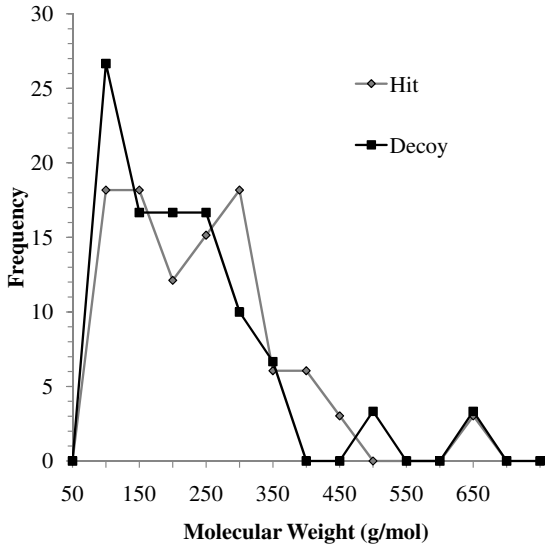


Figure 5.2 Histogram of SlogP for Hits and Decoys

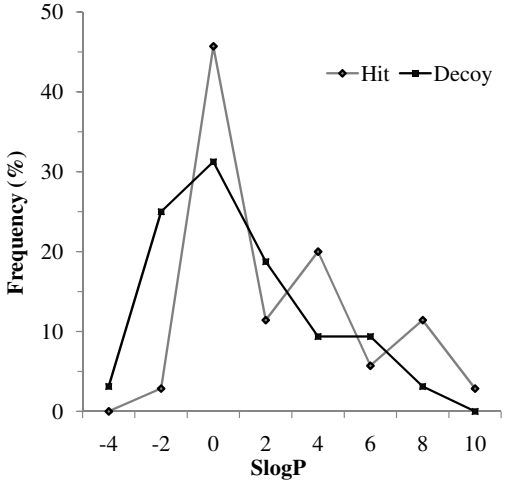


Figure 5.3 Histogram of logS for Hits and Decoys

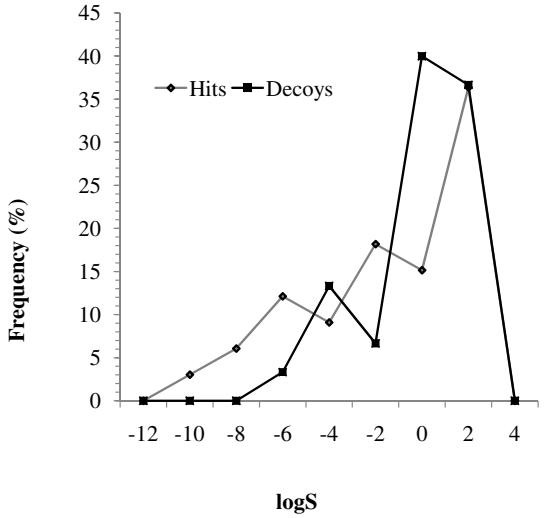


Figure 5.4 Histogram of Buried Surface Area (BSA) for Hits and Decoys

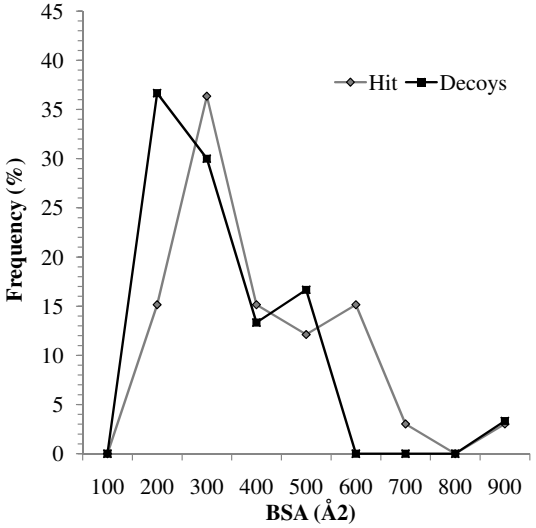
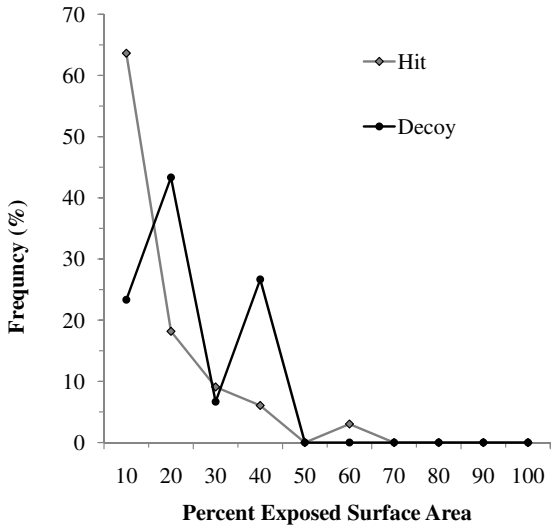


Figure 5.5 Histogram of Percent Exposed Surface Area (%ESA) for Hits and Decoys



5.2.3 Scoring protein-ligand complexes

It is important to stress that all bound ligands, hits and decoys, were maintained in their crystallographic coordinates. This was done to focus on scoring without introducing differences arising form the docking routines. All the protein-ligand structures were prepared in the Chimera software from UCSF [207] by first removing water molecules and metal ions from the complex structures. Next, the protein and ligand were separated for parameter setting, but later reassociated for scoring. Then, hydrogens and charges were added to both the protein and the ligand. The protein atoms were assigned Amber charges, and the ligand was assigned with Gasteiger charges [208]. After preparing the protein and ligand, the binding-energy scores for all the complexes were calculated by using X-Score, AutoDock 4.0, DOCK 4.0, and MDock programs.

5.2.4 Receiver Operating Characteristic Curves

The scoring functions performance was evaluated by comparing the rank ordering of real ligands (true positive hits) versus decoys (false positives) with the receiver operator characteristic (ROC) curves. A perfect scoring method would rank order all hits before decoys, achieving a curve that starts at the origin (0,0), goes straight to the upper left hand corner of the ROC plot (1,0), and then to the upper right (1,1). A scoring method with no predictive power would equally rank hits and decoys, achieving a line starting from the origin (0,0)

going straight to the upper right (1,1). Area under the curve (AUC) provides a quantitative measure for comparison for ROC curves. A perfect scoring function would have a ROC curve with an AUC of 1.0, while the poorer scoring function described above would score an AUC of 0.5, no better than random assignment.

5.3 Results and Discussion

The test set is composed of 33 valid hits and 30 decoys. Both the hit set and the decoy set contain the buffer (4S)-2-methylpentane-2,4-diol (MPD). This compound exemplifies the challenge for scoring functions to be able to distinguish the biological context of binding events. While there are many extraneous molecules that appear in crystal structures, many are not suitable as decoys, especially when restricted to be chemically similar to the hit set of ligands. As Figure 5.6 shows, the sets were carefully chosen so that the two classes could not be easily distinguished by molecular weight, logS, SlogP, or BSA. Similar BSA between the two sets is very important as it is directly proportional to the number of contacts that contribute to the scores.

Figure 5.6 ROC Plot of Percent Exposed Surface Area (%ESA), Exposed Surface Area (ESA), Buried Surface Area (BSA), Molecular Weight (MW), logS, and SlogP. The areas under the curve (AUCs) are noted in the figure legend.

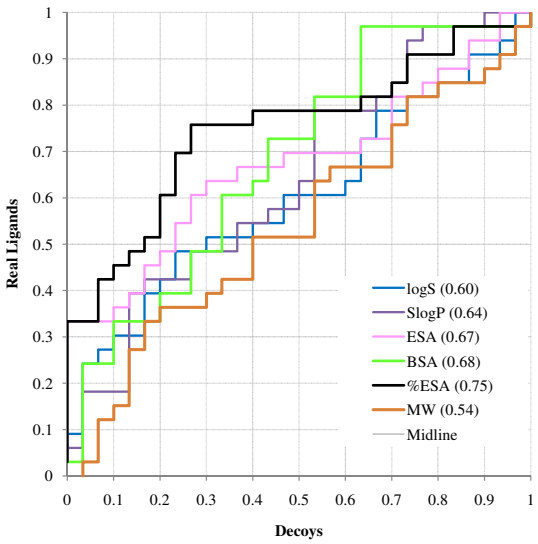


Table 5.4 Number of decoys in the top scoring results.

	In Top 10% (out of 6)	In Top 20% (out of 13)	In Top 40% (out of 25)	In Top 50% (out of 32)
ITScore	0	1	3	7
Dock4	2	2	6	10
AutoDock4	1	2	7	11
X-Score	1	1	5	9

5.3.1 Analysis of Scoring Functions

It is important to note that the scoring with X-Score, DOCK, and ITScore were performed "blindly". As part of our collaboration, the Carlson group provided the Zou lab with the full list of 63 complexes, without noting which were hits and which were decoys. The Zou group scored the full list with and without torsional entropy penalties for the ligands (X-Score, DOCK+tor, ITScore+tor, X-Score-tor, DOCK, ITScore, respectively). ROC plots were generated by the Carlson lab to reveal performance. Rankings with AutoDock (and AutoDock-tor) were performed by the Carlson lab subsequently to include more diversity in the study.

We were delighted that X-Score, AutoDock, DOCK, and ITScore all performed well, preferentially distinguishing hits over decoys (Figures 5.7 - 5.11). This is very promising for extending current scoring functions to new uses in structural proteomics like predicting druggability or function of a protein. Most of the biologically relevant ligands in protein-ligand crystal structures are well buried.[103] Tight-binding ligands have, on average, more BSA than weakly bound ligands.[160] During the creation of the dataset, we were very careful to match the BSA of hits and decoys. The ROC plots show that BSA does not distinguish hits from decoys (Figures 5.6), and this emphasizes that all the scoring functions are out-performing a mere count of contacts in evaluating the ligands.

5.3.2 Torsional Entropy

Hits were ranked over decoys whether or not torsional entropy terms were included, but it is important to note that adding the penalty significantly reduced the false positive rate in the highest-ranked ligands (Figures 5.8 - 5.11). This was seen for all the scoring functions. Including a torsional term in ITScore and DOCK improved performance (AUC increased by 0.02 and 0.05, respectively), and removing the term from X-Score and AutoDock degraded performance (AUC decreased by 0.09 and 0.02, respectively). Of course, removing a term from a scoring function should degrade its performance, but it was interesting that the

changes in AUC are nearly the same for all the scoring functions. Furthermore, the terms are essential to all the scoring functions in the most critical region of the ROC plot (lower left) where the highest-ranked compounds are shown.

A close look at the ranked scores by ITScore with and without the ligand torsional penalty showed that the ranks of two decoy complexes (PDB codes: 8CHO and 1UN8) were significantly changed when the term was included. These decoy complexes are long, chain-like ligands that are not expected to be biologically important (see Figure 5.12). These invalid ligands are large and extremely flexible. For example, the ligand in 8CHO has a total of 18 rotatable bonds, and 1UN8 has 25. In these cases, the ligand's large size results in a good score because of the many contacts, and the penalty term is needed to incorporate the high conformational entropy loss for the ligand when the torsions become restricted.

It was a little serendipitous that the torsional penalty could be added to ITScore and DOCK for this application (ITScore+tor and DOCK+tor, Figures 5.8, 5.9). It is promising that performance could be further improved by properly fitting these terms into DOCK and ITScore. While knowledge-based potentials like ITScore aim to represent all physical contributions to binding, it is still restricted to any limitations in its training set. ITScore's training set is much larger than others, but torsional entropic penalties of the ligand will not be well accounted for unless the training set includes ligands with many rotatable bonds. Pair-wise potentials are iteratively trained by identifying native poses over incorrect poses, but docking ligands with many rotatable bonds is inherently difficult because of their large conformational space. This incompatibility means that pair-wise potentials simply cannot account for this penalty well at this time. Most likely, the most appropriate approach is to iteratively fit a new term as a corrective measure when training the pair-potential.

5.3.3 Top-Scoring Complexes

A set of hits was consistently ranked highly by the scoring functions, see Table 5.5. These hits are diverse, ranging from carbohydrates (1RQ5), steroids (1HDC), arachidonic acid (1VYG), to analogs of retinol (1FEN) and ornithine (1OTH). Most of the hits are substrate or product analogs. One hit is an actual substrate: cellotetrose is bound to the catalytically inactive E796Q cellobiohydrolase mutant (1RQ5). Other hits are nonenzymes binding their ligands. Nonenzyme examples include spermidine/putrescine-binding protein binding spermidine (1POT), β-lactoglobulin (1B0O), and brain fatty acid binding protein (1FDQ). However, each of these hits are biologically relevant and make important contacts with the protein - whether it be in an active site or a dimerization interface.

The bisubstrate analog N-phosphonacetyl-L-ornithine binds in the ornithine transcar-

bamolyase (1OTH) active site, located in a cleft between domains with an IC_{50}=100nM.[209] The principal protein residues that interact with ligand are Asn 199, Asp 263, Ser 267, and Met 268. Asn 199 is involved in domain closure; the catalytic Cys 303 forms a charge relay system with Asp 263 and interacts with the α-amino group of L-ornithine.

The 1POT structure details spermidine binding to spermidine/putrescine-binding protein PotD, with a K_d3.2=μM.[210] The polyamine spermidine binds in the acidic substrate-binding site, located between the N- and C-domains. The 1.8-Å resolution structure clearly shows the oxygen atoms from Thr 35, Glu 36, Tyr 85, Asp 168, Glu 171, Asp 257, and Gln 327 interacting with the three nitrogens of spermidine. The protein formed a hydrogen bond with the terminal amino group of the aminobutyl moiety of the spermidine and formed two hydrogen bonds to the hydroxyl group of Thr 35 and the main-chain carbonyl oxygen of Ser 211.

1FEN is bovine plasma retinol-binding protein (RBP), which was crystalized with the retinol analog axerophtene in a study that examined how RBP recognized different retinol analogs.[211] Axerophtene exhibited the same mode of binding as retinol, binding in the RBP β barrel. The axerophtene analog did not induce conformational changes in a flexible loop region at the entrance of the beta-barrel (although such loop changes were observed with other retinol analogs such as fenretinide and retinoic acid). However, exerophtene's hydrogen atom in place of retinol's hydroxyl end group was cited as being responsible for the reduced affinity and activity in RBP.

The 1CGZ structure of chalcone synthase contains a bound molecule of resveratrol, a product analog.[212] Binding site residues Ser 133, Glu 192, Thr 194, Thr 197 and Ser 338 surround the coumaroyl-derived portion of resveratrol molecule and interact primarily through van der Waals contacts. One prominent hydrogen is formed between the carbonyl oxygen of Gly 216 and the phenolic oxygen of resveratrol.

2ACO structure is a functional dimer of the bacterial outer-membrane lipocalin Blc.[213] Blc is bound with vaccenic acid, an unsaturated C18 fatty acid, in a binding site that spans across the Blc dimer (vaccenic acid covers 89 $Å^2$ and 171 $Å^2$ of the two monomers). Blc binds vaccenic acid with a K_d=2.5μM and displays a preference for lysophospholipids over other fatty acids or phospholipids, suggesting that Blc may be involved in cell envelope lysophospholipid transport.

The 1VYG X-ray crystal structure of fatty acid binding protein Sm14 is used to explain ligand selectivity.[214] Sm14 has numerous tight specific interactions, among which the most important one is a strong directional π-cation interaction between the guanidinium group of Arg 78 and the C8-C9 double bond in arachadonic acid. Sm14 binds arachadonic acid with a K_d=10nM.[214]

1FJ4, the β-ketoacyl-acyl carrier protein synthase is inhibited by thiolactomycin (TLM) with an IC_{50}=25μM.[215] TLM mimics malonyl-ACP and forms strong hydrogen bond interactions with the two catalytic histidines, His 298 and His 333. An unsaturated alkyl side chain in TLM interacts with a small hydrophobic pocket is stabilized by π-stacking interactions that come from intercalating the isoprenoid tail into the space between Pro 272 and its associated peptide bond and the peptide bond between Gly 391 and Phe 392. TLM forms hydrogen bonds with the two active site histidines, His 298 and His 333, and to a network of waters which is held in place by the carbonyl oxygen of Val 270 and by the amine group of Gly 305.

Examining well-scored decoys can be very enlightening, as it can help reveal any caveats and issues of a given scoring function. Furthermore, it is desirable to have "failures" where decoys in true active sites score well because they are chemically similar to the biologically relevant ligand. Of the 30 decoys, 11 were observed to bind in active sites (1D5R, 1D6F, 1D7H, 1N2F, 1PPA, 1Q44, 1QST, 1S2U, 1ZR3, 229L, 2CJP).

Each of the scoring functions highly ranked a similar subset of the decoys, listed in Table 5.6. The scoring functions tended to rank some similar compounds at the top. While there is a general similarity of ranked results, the order and composition varies between scoring functions. There is not just one class that dominates the well scored decoys. Top-scoring decoys include phospholipids that were purified along with the protein (phosphatidyl-glycerol and a modified palmitate) as well as detergent additives (N-octanoyl-sucrose and β-octylglucoside), buffers (TAPS, HEPES, Tris buffers), as well as small organic compounds (guanidinium, cyclohexylammonium, toluene).

There is one structure that was highly ranked by each of the scoring functions. 1D5R, pten tumor suppressor bound to a molecule of tartarate buffer. Here, tartarate binds in the active site, making similar contacts to those seen binding the substrate inositol (1,3,4,5)-tetrakisphosphate.[216] Complexes that were highly ranked by 3 of the scoring functions include 1RJM, 1IZ2, 1LIH, 1PPA, 1Q61, 1RJM, 1UN8, 1ZR3, and 1IZ2. In the 1RJM structure, EP1 is not in the active site of the protein, but rather in a hole in the center of the MenB trimer.[217] The molecules negatively charged sulfonic acid head group interacts via well ordered water molecules with the positively charged Arg 202 side chains. Also among the top scoring decoys is alpha1-antitrypsin (1IZ2), where the the ligand bound is N-octanoyl-sucrose (SUM), a detergent additive in the crystallization matrix. The detergent molecules tail was partially inserted into the protein, not in the active site. The extracellular, ligand-binding domain of the aspartate receptor is bound to 1,10-phenanthroline, a metallopeptidase inhibitor. Phenanthroline is bound to the ligand-binding domain that would physiologically be embedded in the bacterial extracellular membrane, and is known not to

interfere with aspartate binding.[218]

Several top-scoring decoys are in ligand binding sites, reflecting interactions that could be important for substrate binding. For example, in 1QST is a HEPES buffer is bound in the acetyl-CoA binding site of tetrahymena GCN5, a nuclear histone acetyltransferase enzymes.[219] The buffer 2-(N-morpholino)-ethanesulfonic acid (MES) is found in the binding site of macro-domain of human core histone variant macroH2A1.1 (1ZR3).[220] In the steroid sulfotransferase 1Q44, a molecule of the malate buffer binds in the active site (binding pocket is determined by comparing 1Q44 with the homolog 1J99). The bacterial hydroperoxide resistance protein Ohr (1N2F) has two cysteines in the active site Cys 60 and Cys 124 which bind DTT.[221] While these decoys are opportunistic binders, the fact that they score well is encouraging because they are binding in biologically critical sites.

Some of the decoys are identified by modeling molecules into unaccounted electron density based on size, shape, and chemical environment. Usually the molecule is readily identified as a component of the crystallization matrix. In δ5-3-ketosteroid isomerase(8CHO), a molecule of the polyethylene glycol was modeled to fit the density.[222] Other cases exist where the ligand was not from the crystallization matrix. Analine (ANL) is seen to bind near a proposed phospholipid substrate site in the crystal structure of phospholipase A2 (1PPA). Its origin, however, is a mystery and was identified by fitting electron density.[223] Occasionally, the the molecule is assumed to have been co-purified along with the protein. In the dihyroxyacetone kinase structure (1UN8), the ligand is identified as 2-myristic(C14)-3-palmitic(C16)-phospholipid.[224] Here, the lipid binds the protein in ellipsoidal shape pocket (5 by 11 Å wide), two acyl chains extend 15 Å into the pocket where they are surrounded by apolar side chains, and the lipid head group lies exposed to solvent at the entrance of the pocket.

5.4 Conclusion

This study puts forth a new criteria for evaluating scoring functions: the ability to discern between opportunistic binding by opportunistic ligands (decoys) and biologically important ligands (hits). Accordingly, a new test set is put forth, containing 33 hits and 30 decoy structures. The decoy and hits structures show similar distributions of physicochemical properties such as MW, hydrophobicity, solubility, and BSA. Four different scoring functions, representative of knowledge-based, force-field, and empirical functions, are used to evaluate this test set. The results show that these four scoring functions are able to discern decoys from hits and achieve ROCs with AUC of 0.85 (ITScore+tor), 0.72 (DOCK4+tor),

Figure 5.7 ROC Plot of scoring functions, BSA and ESA (optimal performance for ITScore and DOCK4 included torsional penalties

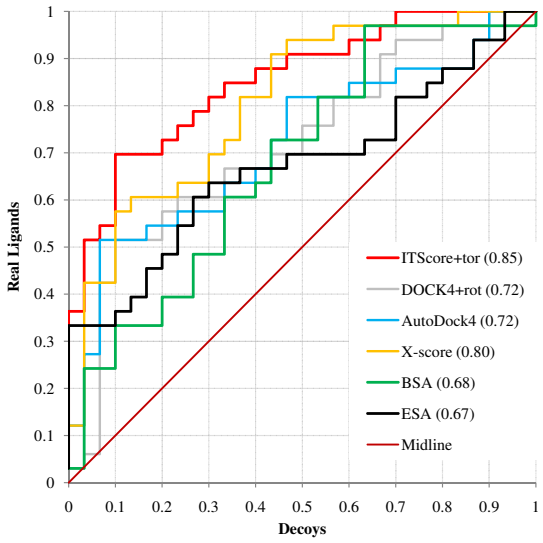


Figure 5.8 ROC Plot of ITScore

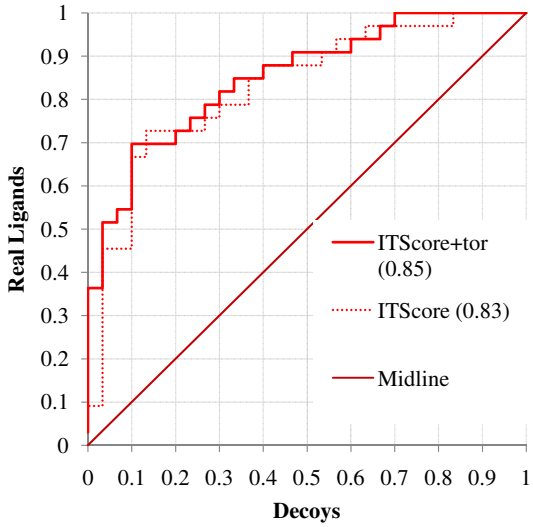


Figure 5.9 ROC Plot of DOCK4

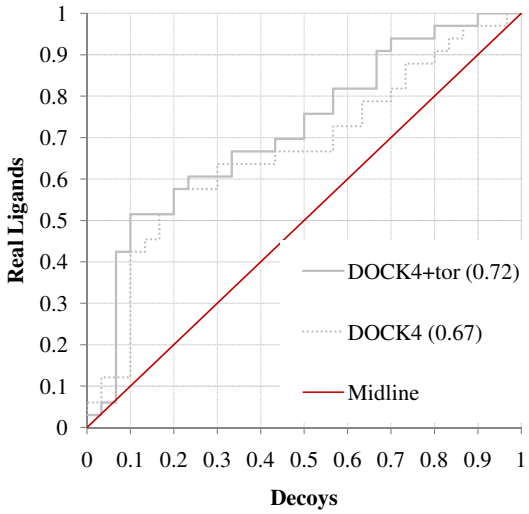


Figure 5.10 ROC Plot of AutoDock4

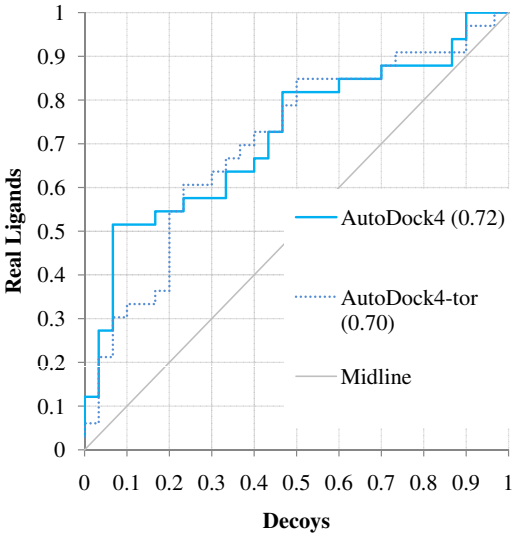


Table 5.5 Best Scoring Hits. Hits ranked high by two or more scoring functions are in plain text. Unique complexes are in italics.

ITScore+tor				**DOCK4+tor**			
PDB	Ligand	Rank	%ESA	PDB	Ligand	Rank	%ESA
1RQ5	CTT	1	17.6	1OTH	PAO	1	1.9
1OTH	PAO	2	1.9	1POT	SPD	3	0.0
1POT	SPD	3	0.0	1RQ5	CTT	5	17.6
1W3J	OXZ	4	5.2	*2CCS*	*4BH*	*6*	*6.6*
1CHM	*CMS*	*5*	*0.8*	1HDC	CBO	7	23.7
1T0O	*GAL*	6	*4.6*	1CGZ	STL	8	1.6
1PX8	*XYP*	*7*	*5.1*	1FEN	AZE	9	11.2
1FEN	AZE	8	6.4	1FJ4	TLM	10	3.7
2ACO	VCA	9	12.1	1VYG	ACD	11	6.5
1VYG	ACD	10	6.5	1UTJ	ABN	12	10.4

AutoDock4				**X-Score**			
PDB	Ligand	Rank	%ESA	PDB	Ligand	Rank	%ESA
1OTH	PAO	1	1.9	1HDC	CBO	1	23.7
1POT	SPD	2	0.0	1RQ5	CTT	2	17.6
1CGZ	STL	3	1.6	*1UN8*	*MYY*	*3*	*15.2*
1VYG	ACD	4	6.5	1CGZ	STL	4	1.6
1FJ4	TLM	6	3.7	1VYG	ACD	5	6.5
2ACO	VCA	7	12.1	1VG0	GER	7	28.1
1UTJ	ABN	8	10.4	1YXX	LI7	8	5.0
1VG0	GER	9	28.1	2ACO	VCA	10	12.1
1YXX	LI7	10	5.0	*1P4F*	*DRG*	*11*	*5.7*
1W3J	OXZ	12	5.2	1FJ4	TLM	12	3.7

0.72 (AutoDock4), and 0.81 (X-Score). The approximation of ligand torsional entropy in the scoring functions is shown to be important in ranking the protein-ligand complexes.

This test set has potential to help improve algorithms used in molecular docking by providing a different measure for docking success. Its further development will be essential to extending scoring functions to new purposes like identifying protein function or estimating the druggability of a pocket.

Table 5.6 Best Scoring Decoys. Unique complexes are in italics, decoys commonly ranked high are in plain text. Acceptable failures of decoys in real binding sites are marked with a star.

ITScore+tor				**DOCK4+tor**			
PDB	Ligand	Rank	%ESA	PDB	Ligand	Rank	%ESA
1UN8	MYY	13	15.2	1RJM	EP1	2	3.6
1D5R	TLA*	19	5.2	1D5R	TLA*	4	5.2
1IZ2	SUM	21	30.1	1ZR3	MES*	17	5.4
2I3A	*BTB*	*27*	*34.6*	*1LI0*	*PHN*	*21*	*39.4*
1ZR3	MES*	28	5.4	1IZ2	SUM	22	30.1
1TTO	TRS	29	32.0	1QST	EPE*	23	13.3
1RJM	EP1	31	3.6	229L	GAI	26	17.3
1LIH	PHN	33	39.4	1TT0	TRS	28	32.O
1APM	OCT	35	17.5	1SHV	MA4	29	35.4
1Q61	MG8	37	18.7	1PPA	ANL*	30	5.2

AutoDock4				**X-Score**			
PDB	Ligand	Rank	%ESA	PDB	Ligand	Rank	%ESA
1D5R	TLA*	5	5.2	1UN8	MYY	5	15.2
1ZR3	MES*	11	5.4	1SHV	MA4	16	35.4
229L	GAI*	20	17.3	1IZ2	SUM	17	30.1
1PPA	ANL*	21	5.2	1RJM	EP1	23	3.6
1UN8	MYY	22	15.2	1Q61	MG8	25	18.7
1D7H	*DMS**	24	13.2	1APM	OCT	26	17.5
1QST	EPE*	25	13.3	*1BF3*	*HTO*	*27*	*3.2*
1Q44	*MLA**	*27*	*6.6*	1PPA	ANL*	29	5.2
1N2F	*DTT**	*28*	*16.8*	1LIH	PHN	30	39.4
1Q61	MG8	29	18.7	1D5R	TLA*	33	5.2

Figure 5.11 ROC Plot of X-Score

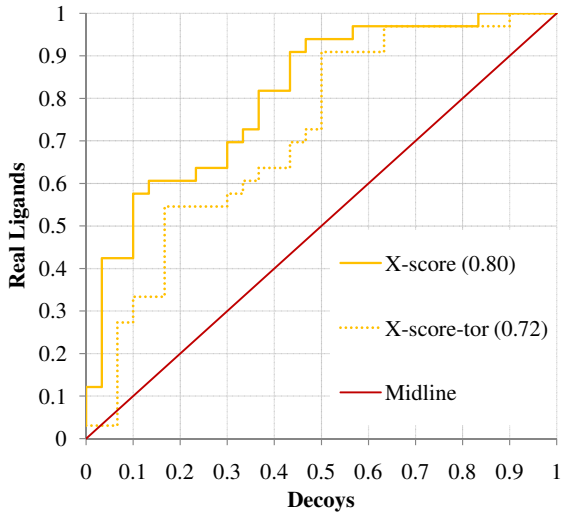
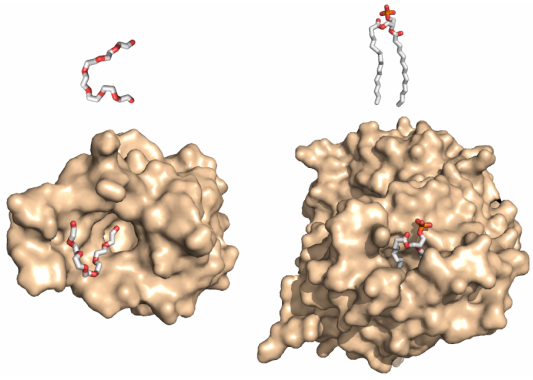


Figure 5.12 The phospholipid ligands and their corresponding complexes for 8CHO (left) and 1UN8 (right). The ligands are represented in stick mode, and the protein is shown by molecular surface. The figure was prepared by PyMOL.[2]



Appendices

Appendix A

BindingMOAD.org Architecture

A.1 Introduction

Communication and dissemination of information play an essential role in science. Until discoveries and insight are shared with others, they remain locked and hidden. The data in Binding MOAD has been made available online since 2004. This section describes the technology and principles of software development to create and disseminate Binding MOAD from a high-level architectural view.

Binding MOAD uses a Java 2 Enterprise Edition (J2EE) framework and 3-tier model (client, application, and database). A framework is a set of classes and interfaces that are designed to help solve a specific problem and act like a skeleton code for a developer to build upon. The 3-tier model is mechanism for describing a separation of function. The user interacts with the application through the client tier. The client tier is comprised of a standard web browser (such as Internet Explorer) which accesses the website over the Internet. The middle tier coordinates the users HTTP requests and handles calculations, logic, and decisions to manage information flow and dynamically generate web pages. The database tier is where data is stored and retrieved.

A.2 Client Tier

The client tier is how the user interacts with the application. In this case, the client tier is comprised of a standard web browser (such as Internet Explorer, Firefox, Safari, etc.) to view Binding MOAD via the Internet at "http://BindingMOAD.org". The web is an ideal mechanism for making Binding MOAD available to the scientific community (as opposed to requiring users to install programs on their own computers or making data available via

ftp). The webserver is publicly accessible, and the user is not required to register or logon to access to the data.

One of the services available via Binding MOAD is the EolasViewer, a Java based program which allows you to view the protein-ligand cavity for a given crystal structure in Binding MOAD. The EolasViewer runs using Webstart, which is integrated into Java. Thus, the user needs to have a working Java environment installed along with the Java 3D API, which is freely available in both proprietary and open source versions. The EolasViewer supercedes the GoCAVViewer for visual examination of protein-ligand complexes. A working Java install is also required for the ChemAxon viewer, which allows the user to view the ligand in 2D.

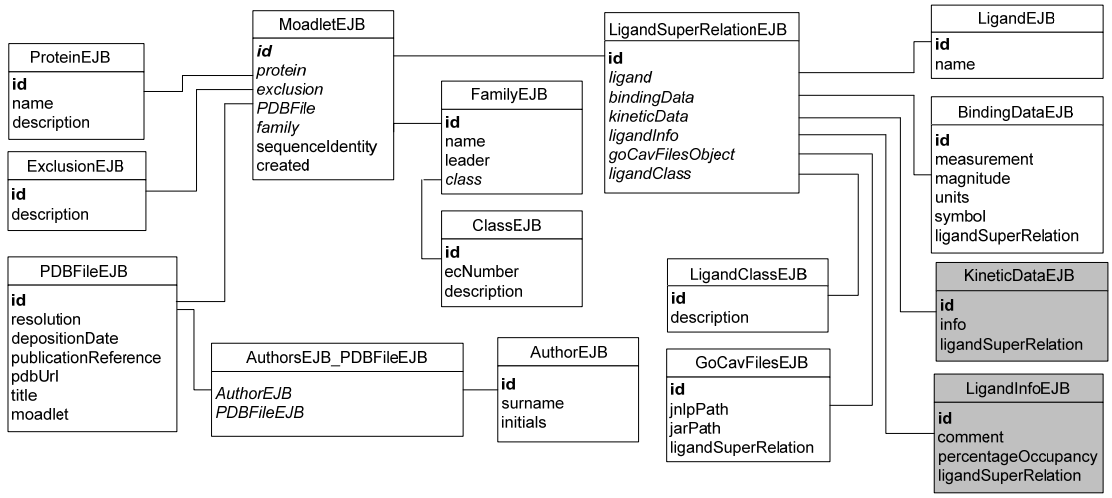
A.3 Database Tier

The database tier is how data is stored and retrieved (see Figure A.1). Here, MySQL is used as the database for persistence. Data from hand curation and data extraction include PDB id, EC class, homolog family, binding affinity data, and classification of each ligand in the entry (ligand class: ligand, multipart ligand, common cofactor, metal ion, covalent modification to the protein, or crystal additive). The data include the reasons any PDB structure was excluded. This is parsed and loaded into the server and is stored in a MySQL database. The data can either be loaded though a java interface through Binding MOAD, or it can be loaded directly into the MySQL database.

A.4 Middle Tier

The middle tier, also called the web tier, is responsible for most of the functionality. It is the workhorse that handles HTTP requests from the client, communicates with the database, and manages the information flow to dynamically generate the webpages. We are running JBoss 4, an open source Java2 Enterprise Edition (J2EE) application server. The JBoss application server provides and integrate several technologies, including a Tomcat web containter, Enterprise Java Beans (EJB) for data management, and Jakarta Struts framework[225, 226, 227]. The Web container manages presentation of HTML, XML, JSP, Servlets, and details relating to application state such as connection-pooling, security, and session handling. The big advantage of JSP is its ease of maintenance during the semi-annual updates; new and corrected data are automatically propagated to the data pages. JBoss compiles JSP to dynamically

Figure A.1 The EJB data model in BindingMOAD. These MySQL tables represent the types of data. Primary keys for each table are listed and foreign keys are given in italics. The data model is organized around two central tables MoadletEJB and LigandSuperRelationEJB. MoadletEJB describes a protein entry in Binding MOAD (PDB id, protein family and class, authors who submitted the structure, etc). LigandSuperRelationEJB describes the relationships between proteins and the ligands (name of the ligand, binding data, valid/invalid ligand, and files needed for the GoCAVviewer). The tables in "gray" (KineticDataEJB and LigandInfoEJB) represent data and features that could be added to Binding MOAD; they are shown to illustrate their place as appropriate data is expanded.



generate valid, standards-compliant XHTML and CSS. There are a variety of convenient tag libraries for efficient JSP development.

The EJB container mediates communication between the database tier and the web container and handles many operation critical behaviors. The data modeling in EJB allows for expansion without significant changes to the code. The Jakarta Struts framework implements the Model-View-Controller (MVC) using servlets and Java Server Pages (JSP) technology. MVC is a design pattern that essentially decouples the business logic in the application from the presentation, which allows the developer to change either the appearance or the business logic, without significantly affecting the other components.[228] A framework can be defined as a reusable abstraction of code wrapped in a well-defined API, akin to a software skeleton that controls program flow of control.

A.5 Maintenance and Expansion

A key indication of good code is maintainability and ease of adding new features. Since Binding MOAD was first release, we have added chemical information for the ligands

(names, formulas, SMILE strings, 2D pictures, etc) and increased the number of searchable fields. Additional changes that can be added include more levels of classification based on different percent similarity (50%, 75%, 100% identical proteins), searching against structural similarities of binding sites (or protein structure as a whole), on-the-fly searching of proteins via BLASTp, adding kinetic data to affinity data, providing a download feature for the user to obtain the structures and binding data from search results, SMILE string searches, and ability to search and calculate chemical similarities of ligands. Adding new data and relationships between data is possible due to a structured and flexible software infrastructure using proven technologies.

Data Integrity

Great care is taken to ensure that the data in Binding MOAD is accurate and correct. In cases where binding data is not available, a link is provided to encourage users to deposit information. Our deposition page contains fields for binding data, its reference in the literature, and the user's name and e-mail address. While other databases provides pages where users actively add data to their existing database, we are concerned with security and data integrity, so we opt to review and analyze all data before it is entered into Binding MOAD.

To ensure that the data integrity is maintained on the server, we have adopted several security practices including a hardened server (running a stripped-down system without unnecessary services), have consciously taken a number of precautions to aid security such as not displaying unfiltered user-entered text (to prevent insertion-type attacks), placing JSP files behind WEB-INF (to prevented inappropriate access and ensures that the user views work). The JBoss web server provides many security enhancements as part of its adherence to the J2EE specifications that explicitly forbid many common types of remote-initiated attacks.

A.6 Conclusion

These technologies provide a robust and scalable infrastructure with a large community appropriate for this project. It allows seamless addition of new features and data. This modular software infrastructure allows for individual components to be easily adapted and expanded to allow additional plugins, relationships, and data.

Appendix B

Updating Binding MOAD - Data Management and Information Workflow

B.1 Introduction of Protein-Ligand Databases

In medicinal chemistry and structure-based drug design, a major aim is to discover small molecules that bind a target protein with tight affinity, but predicting protein-ligand interactions is not trivial. Proteins are dynamic, and some can bind a wide variety of ligands.[11] Ligands can also change conformation upon binding. Much work has gone into developing docking, scoring functions, and estimating binding affinities.[3] These types of studies rely on accurate databases of protein-ligand complexes which are used to train models and fit equations. Such databases are also essential to data-mining studies used to derive the biophysical patterns that dictate ligand binding.[116]

Currently, there are several available protein-ligand databases, such as Binding MOAD[115], LPDB[30], MSDsite[66], Relibase[99], BindingDB[47], PDBbind[69], eF-Site[51], PDB-Ligand[76], SuperLigands[95], PLD[78], HET-PDB[229], PDBsite[230], Ligand Depot[98], sc-PDB[87], AffinDB[44], KiBank[64], and PDBLig[74] , each having a different scope. Some simply list compounds from the Protein Data Bank (PDB)[61], some focus solely on structural analyses, some present only binding data, and others are cohesive datasets with all of these elements.

Binding MOAD is the largest collection of curated protein structures with biologically relevant ligands annotated with binding affinities from the literature.[116, 115] Our dataset distinguishes itself from the gamut of protein-ligand databases because of its extensive number of entries, high quality hand curation, and regular addition of data. Binding MOAD contains all appropriate complexes (protein-ligand, protein-ligand-cofactor, and protein-cofactor), whether or not binding data is available, making it four times larger than its closest

competitor.

B.2 Updating the Protein-Ligand Structures in Binding MOAD

Maintaining the online resource, correcting errors, and keeping pace with the growth of the PDB has well exceeded the initial efforts to create Binding MOAD from 2001-2003.[115]

There are several phases to our semi-annual update process. First, any structures identified as incorrect are removed from the current dataset; these are structures added to the PDB's "obsolete list" since the last update. Second, we identify which new PDB entries meet our criteria. Third, we asses the literature. And lastly, the new entries are added into the existing database and published on the web. Removing obsolete entries and adding new entries is relatively straightforward. Our discussions below focus on the identification of new structures and their annotation with information from the literature.

It is straightforward to obtain the subset of new structures in the PDB, added subsequent to our last update. These entries are then evaluated through a series of Perl scripts to ensure that they meet our requirements: only X-ray crystal structures of resolution greater than 2.5Å (no NMR structures), must contain a protein but no nucleic acid macromolecules, must contain at least one valid ligand.[115]

The most complicated aspect is the evaluation of ligands. Only non-covalent ligands are considered, and these are identified by calculating the minimum distance between the protein and the molecule to ensure that it is longer than a covalent bond. No crystallographic additives are considered valid, such as buffers (e.g. Tris, CHAPS), ions (phosphate, chloride), solvents (water, DMSO, acetone), detergents (Triton-X, polyethylene glycol), metal ions (Mg^{2+}), or catalytic centers that are part of the protein (4Fe-4S cluster). Small nucleic acid chains (4 nucleic acids or less) or peptides (10 amino acids or less) are valid ligands by our definitions.

All structures are manually viewed to verify the validity of the ligands. This is essential when the minimum distance between the ligand and protein is longer than a traditional bond but too short for typical van der Waals contact. The complex is also verified by examining the paper which reports the structure. Visualization and the original reference are particularly needed in the case of certain molecules that we label "suspect"; these are molecules like citrate which can be a valid ligand but can also be a buffer component.

B.3 Annotating the Structures with Information from the Literature

Tremendous effort has gone into annotating each entry with experimental data from the literature. Each structure is augmented with information on its functional relationships to other structures in the database, classification of all ligands in the structure (valid/invalid/covalently attached), and binding data. For valid ligands, we also note if any atoms are unresolved or resolved in multiple orientations because these cases may be problematic for certain applications.

Every PDB file contains the primary reference in the literature. These are read to assess the quality of the structure, the validity of the ligands, and the binding data. Full-length HTML and PDF formatted papers are obtained from the various publishers. Assessing the literature is a sizable task, covering $\sim$ 2000 papers each year! An improvement was needed to make this exercise more tractable: BUDA (Binding Unstructured Data Analysis) makes the curation more efficient and accurate. It consists of two primary applications integrated using a bibliographic control. The first is a text processing application built upon the GATE3 platform, provided by Sheffield University. The second is a custom curation workflow tool developed using the Ruby programming language.

B.3.1 Natural Language Processing in BUDA

Natural Language Processing (NLP) refers to a broad category of techniques used to analyze and derive understanding from unstructured information, primarily human-readable text. A subset of these techniques, the identification and assembly of desired information elements from within a document or set of documents, is commonly referred to as information extraction (IE). IE for biochemical data from the scientific literature is a difficult and heavily-researched area.[231, 232] For Binding MOAD's curation, we identify precise quantitative information related to the interactions between specific proteins and ligands in biomedical experiments.

The NLP in BUDA begins with the evaluation of full text articles using GATE (Figure B.1). GATE processes text to create annotations that can be stored as data for an application in XML and/or outputted into a modified version of the input document (Figure B.1). The standard distribution of GATE includes several processing components, the most important of which are the ANNIE plug-ins.

ANNIE includes tokenisers, a sentence splitter, a part-of-speech tagger, and ANNIE Gazetteer, a plug-in designed for term lookup. [114] Each of these creates annotations of its

Figure B.1 BUDA's GATE pipeline consists of ANNIE plug-ins, a set of modified lookup lists for the Gazetteer, a cascade of eight JAPE grammars, and final processing and export tools. Our additions to the lookup lists consist of keywords, constant names, molar unit symbols, etc. The JAPE transducers recombine the annotations created by ANNIE and the modified Gazetteer to annotate larger phrases and full sentences. For instance, one transducer is used to group cardinal numbers with molar units (e.g., nM, mM, pM, etc.) and annotate the groups as BUDAUnits. A second transducer then identifies and highlights patterns where a constant name is very near a BUDAUnit. This cascade forms an annotation that is a very strong predictor for binding data.

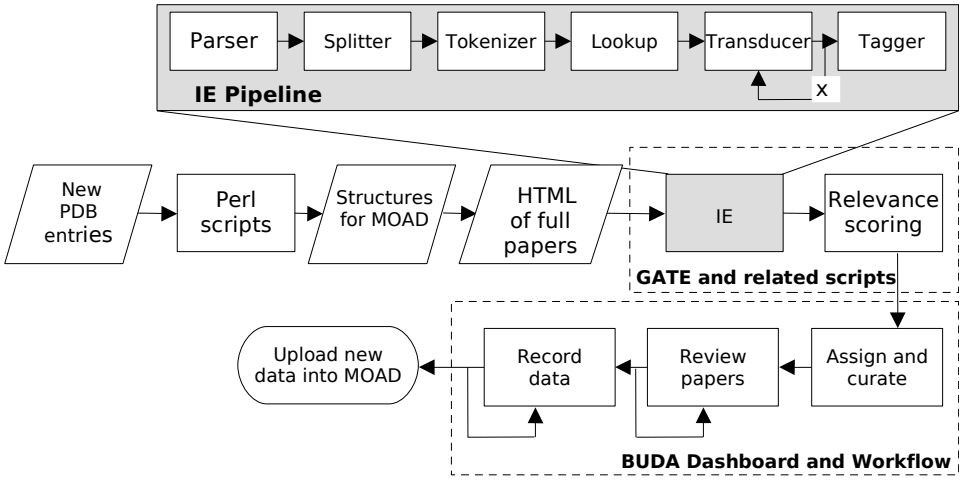
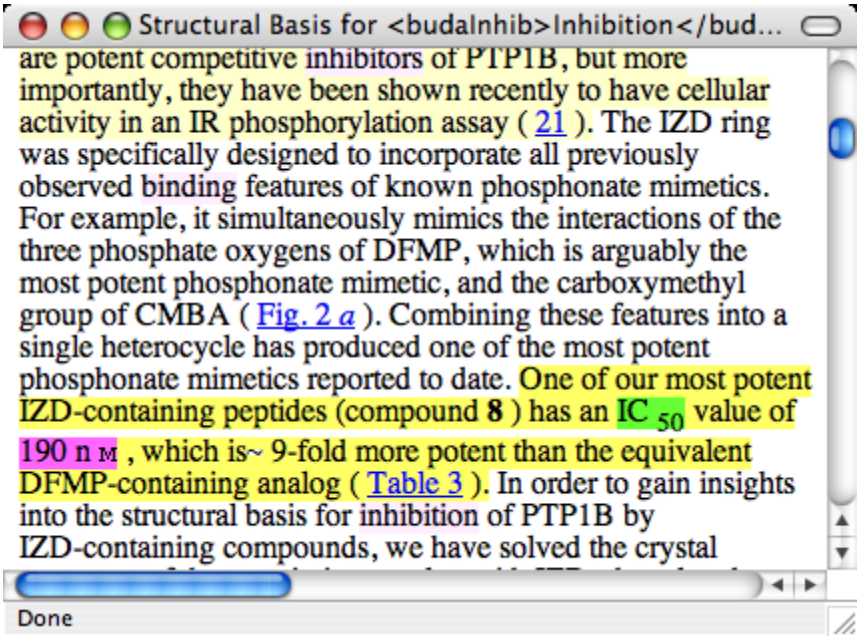


Figure B.2 Markup of the NLP on the HTML of a representative article. Information is highlighted in the text as well as the figure captions. Markup is also highlighted in tables (not shown)



own specific type relating to the features of a parsed document. GATE's JAPE transducer provides grammar processing. This allows us to create our own pattern-recognition rules and controls. The rules can apply additional annotations or conditionally trigger GATE functionalities using Java code.

Finally, GATE provides an overall framework, including a run-time user interface and a Java API, in which these components can be run in a controlled "pipeline." In our case, we specifically developed the pipeline to process scientific articles as HTML files as they are loaded into GATE corpora. (PDF files can also be processed, but with less accuracy and functionality.)

B.3.2 Information Extraction and Information Retrieval in BUDA

Many previously-published IE projects extract values only where they are easily accessible via a computational (text mining) tool[233, 94, 234], but the Binding MOAD project seeks to identify binding values from all available articles. The values may appear within sentences of the text, entries of tables, and even within figures (Figure B.2). Data outside the general text will usually be missed with typical IE applications. However, many indirect references to binding data can be identified and extracted from the full text. These indirect references can be used to create a relevance score, indicating the likelihood of the article containing desired binding data. The creation of such a score is commonly used in information retrieval (IR)[235], a second branch of NLP. Thus, we have designed BUDA to process text and use a combined application of both IE and IR.

The appearance of many related terms in meaningful combinations within sentences will cause the passage to be heavily annotated within BUDA. From an IE perspective, the annotations are useful to the curators reviewing the article because they highlight the probable location of the appropriate data (Figure B.2). From an IR perspective, the combination of terms will serve to increase the relevancy score for the article as a whole. Higher scores identify articles that are the most likely to contain binding data, and lower scores correspond to articles that most likely do not contain binding data.

We developed a separate module using Ruby to perform BUDA's relevance scoring. The module performs operations on the XML files created by GATE's NLP pipeline processing. These XML files contain annotations of different types that are assessed and summarized by the module. The relevance score is a weighted sum of eight parameters "including different phrases and specific key words" that are tallied by the application. With each parameter, the number of occurrences is weighted by a distinct multiplicative factor, and the results are summed to obtain a final score for each article. Thus, the score in the BUDA

application is completely determined by the choice of parameters and weighting factors. The weighting factors were largely determined by inspection. Using a test data set, in which the presence or absence of binding data was known, the weighting factors were modified until the algorithm's determination of the presence of binding data performed optimally. The overall algorithm also uses the length of the target article for scaling purposes.

The relevance scores created in BUDA are exported to the curation application as XML with the annotated originals in HTML. Curators then use the scores to sort articles and focus on those with the best probability of containing binding data.

B.4 Result of Updates

After our 2006 update of Binding MOAD, it contained all appropriate entries deposited within the PDB prior to 1/1/2007. Binding MOAD contained 9837 protein-ligand complexes organized into 3151 unique protein families with 4665 unique ligands. The workflow described above was used to assess the crystallography papers for all 9837 structures and obtain binding data. Binding data was available for 2950 (30%) of the structures. The tool was used again in the 2007 update to provide the current version of Binding mOAD, which now exceeds 10,000 structures.

These numbers express our dedication to constantly improving Binding MOAD and likewise reflect on the wealth Binding MOAD has to offer. Each annual update has added $\sim$ 1500 protein-ligand structures, with $\sim$ 500 having binding data. As the PDB continues its significant expansion, new technologies will be even more crucial for adapting for curating and annotating this resource with knowledge from the literature.

Bibliography

[1] Gunasekaran, K. and Nussinov, R. (2007) How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. *J Mol Biol 365*, 257–73.

[2] DeLano, W. (2002) The pymol molecular graphics system (DeLano Scientific, San Carlos, CA, USA.).

[3] 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.

[4] Warren, G. L., Andrews, C. W., Capelli, A.-M., Clarke, B., LaLonde, J., Lambert, M. H., Lindvall, M., Nevins, N., Semus, S. F., Senger, S., Tedesco, G., Wall, I. D., Woolven, J. M., Peishoff, C. E., and Head, M. S. (2006) A critical assessment of docking programs and scoring functions. *J Med Chem 49*, 5912–5931.

[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] 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.

[7] Fischer, E. (1894) . *Ber. Dtsch. Chem. Ges. 27*, 2985–2993.

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

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

[10] Jr, D. E. K., Jr, W. J. R., and Erwin, M. J. (1958) Protein structure and enzyme action. *Fed Proc 17*, 1145–1150.

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

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

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

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

[15] 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.

[16] Volkman, B. F., Lipson, D., Wemmer, D. E., and Kern, D. (2001) Two-state allosteric behavior in a single-domain signaling protein. *Science 291*, 2429–2433.

[17] Carlson, H. A. (2002) Protein flexibility is an important component of structure-based drug discovery. *Curr Pharm Des 8*, 1571–1578.

[18] Heringa, J. and Argos, P. (1999) Strain in protein structures as viewed through nonrotameric side chains: I. their position and interaction. *Proteins 37*, 30–43.

[19] Heringa, J. and Argos, P. (1999) Strain in protein structures as viewed through nonrotameric side chains: II. effects upon ligand binding. *Proteins 37*, 44–55.

[20] Hilser, V. J., Dowdy, D., Oas, T. G., and Freire, E. (1998) The structural distribution of cooperative interactions in proteins: analysis of the native state ensemble. *Proc Natl Acad Sci U S A 95*, 9903–9908.

[21] Shoichet, B. K., Baase, W. A., Kuroki, R., and Matthews, B. W. (1995) A relationship between protein stability and protein function. *Proc Natl Acad Sci U S A 92*, 452–456.

[22] Keskin, O., Ma, B., and Nussinov, R. (2005) Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol 345*, 1281–94.

[23] Halperin, I., Wolfson, H., and Nussinov, R. (2004) Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure 12*, 1027–38.

[24] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comp Chem 4*, 187–217.

[25] Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J., and Kollman, P. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society 117*, 5179–5197.

[26] Meng, E. C., Shoichet, B. K., and Kuntz, I. D. (1992) Automated docking with grid-based energy approach to macromolecule-ligand interactions. *J. Comput. Chem. 13*, 505–524.

[27] Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003) Improved protein-ligand docking using GOLD. *Proteins 52*, 609–623.

[28] Morris, G., Goodsell, D., Halliday, R., Huey, R., Hart, W., Belew, R., and Olson, A. (1998) Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry 19*, 1639–1662.

[29] Ewing, T. J., Makino, S., Skillman, A. G., and Kuntz, I. D. (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des 15*, 411–428.

[30] 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.

[31] Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov 3*, 935–949.

[32] Bhm, 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.

[33] Bhm, H. J. (1998) Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J Comput Aided Mol Des 12*, 309–323.

[34] Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996) A fast flexible docking method using an incremental construction algorithm. *J Mol Biol 261*, 470–89.

[35] Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., and Mee, R. P. (1997) Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des 11*, 425–445.

[36] Cozzini, P., Fornabaio, M., Marabotti, A., Abraham, D. J., Kellogg, G. E., and Mozzarelli, A. (2002) Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 1. models without explicit constrained water. *J Med Chem 45*, 2469–2483.

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

[38] Huang, S.-Y. and Zou, X. (2006) An iterative knowledge-based scoring function to predict protein-ligand interactions: I. derivation of interaction potentials. *J Comput Chem 27*, 1866–1875.

[39] Muegge, I. and Martin, Y. C. (1999) A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem 42*, 791–804.

[40] Gohlke, H., Hendlich, M., and Klebe, G. (2000) Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol 295*, 337–356.

[41] DeWitte, R. and Shakhnovich, E. (1996) SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. methodology and supporting evidence. *J.Am.Chem.Soc. 118*, 11733–11744.

[42] Huang, S.-Y. and Zou, X. (2006) An iterative knowledge-based scoring function to predict protein-ligand interactions: II. validation of the scoring function. *J Comput Chem 27*, 1876–1882.

[43] Feher, M. (2006) Consensus scoring for protein-ligand interactions. *Drug Discov Today 11*, 421–428.

[44] 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*, D522–6.

[45] Dani, M., Manca, F., and Rialdi, G. (1981) Calorimetric study of concanavalin a binding to saccharides. *Biochim Biophys Acta 667*, 108–117.

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

[47] Chen, X., Liu, M., and Gilson, M. (2001) BindingDB: A web-accessible molecular recognition database. *Combinatorial Chemistry & High Throughput Screening 4*, 719–725.

[48] Liu, T., Lin, Y., Wen, X., Jorissen, R. N., and Gilson, M. K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res 35*, D198–201.

[49] Chen, X., Ji, Z. L., Zhi, D. G., and Chen, Y. Z. (2002) CLiBE: a database of computed ligand binding energy for ligand-receptor complexes. *Comput Chem 26*, 661–6.

[50] Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res 34*, D668–72.

[51] Kinoshita, K., Furui, J., and Nakamura, H. (2002) Identification of protein functions from a molecular surface database, ef-site. *J.Struct.Funct.Genomics 2*, 9–22.

[52] Kinoshita, K. and Nakamura, H. (2003) Identification of protein biochemical functions by similarity search using the molecular surface database ef-site. *Protein Sci 12*, 1589–1595.

[53] Kinoshita, K. and Nakamura, H. (2005) Identification of the ligand binding sites on the molecular surface of proteins. *Protein Sci 14*, 711–8.

[54] Standley, D. M., Kinjo, A. R., Kinoshita, K., and Nakamura, H. (2008) Protein structure databases with new web services for structural biology and biomedical research. *Brief Bioinform 9*, 276–285.

[55] Kinoshita, K., Murakami, Y., and Nakamura, H. (2007) eF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape. *Nucleic Acids Res 35*, W398–W402.

[56] Lopez, G., Valencia, A., and Tress, M. (2007) FireDB–a database of functionally important residues from proteins of known structure. *Nucleic Acids Res 35*, D219–D223.

[57] Porter, C. T., Bartlett, G. J., and Thornton, J. M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res 32*, D129–D133.

[58] Lpez, G., Valencia, A., and Tress, M. L. (2007) firestar–prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res 35*, W573–W577.

[59] Zemla, A. (2003) LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res 31*, 3370–3374.

[60] Biswas, D., Roy, S., and Sen, S. (2006) A simple approach for indexing the oral druglikeness of a compound: discriminating druglike compounds from nondruglike ones. *J Chem Inf Model 46*, 1394–1401.

[61] 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.

[62] Gardner, S. and Thornton, J. (1998) Iditis: protein structure database. *Acta Crystallogr D Biol Crystallogr 54*, 1071–7.

[63] Ji, Z. L., Chen, X., Zhen, C. J., Yao, L. X., Han, L. Y., Yeo, W. K., Chung, P. C., Puy, H. S., Tay, Y. T., Muhammad, A., and Chen, Y. Z. (2003) KDBI: Kinetic Data of Bio-molecular Interactions database. *Nucleic Acids Res 31*, 255–7.

[64] Zhang, J., Aizawa, M., Amari, S., Iwasawa, Y., Nakano, T., and Nakata, K. (2004) Development of KiBank, a database supporting structure-based drug design. *Comput. Biol. Chem. 28*, 401–407.

[65] Dessailly, B. H., Lensink, M. F., Orengo, C. A., and Wodak, S. J. (2008) Ligasite–a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res 36*, D667–D673.

[66] 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.

[67] Golovin, A. and Henrick, K. (2008) Msdmotif: exploring protein sites and motifs. *BMC Bioinformatics 9*, 312.

[68] 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.

[69] 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.

[70] Wang, R., Lu, Y., Fang, X., and Wang, S. (2004) An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J Chem Inf Comput Sci 44*, 2114–25.

[71] Wang, R., Lu, Y., and Wang, S. (2003) Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem 46*, 2287–2303.

[72] 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.

[73] 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.

[74] Chalk, A. J., Worth, C. L., Overington, J. P., and Chan, A. W. E. (2004) PDBLIG: Classification of small molecular protein binding in the Protein Data Bank. *J.Med.Chem. 47*, 3807–3816.

[75] 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.

[76] 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.

[77] Kabsch, W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr A 32*, 922–923.

[78] 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.

[79] Prabakaran, P., An, J., Gromiha, M. M., Selvaraj, S., Uedaira, H., Kono, H., and Sarai, A. (2001) Thermodynamic database for protein-nucleic acid interactions (ProNIT). *Bioinformatics 17*, 1027–1034.

[80] Kumar, M. D. S., Bava, K. A., Gromiha, M. M., Prabakaran, P., Kitajima, K., Uedaira, H., and Sarai, A. (2006) Protherm and pronit: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res 34*, D204–D206.

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

[82] Schmitt, S., Kuhn, D., and Klebe, G. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol 323*, 387–406.

[83] 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.

[84] Gunther, J., Bergner, A., Hendlich, M., and Klebe, G. (2003) Utilising structural knowledge in drug design strategies: applications using Relibase. *J Mol Biol 326*, 621–636.

[85] Kuhn, D., Weskamp, N., Schmitt, S., Hllermeier, E., and Klebe, G. (2006) From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J Mol Biol 359*, 1023–1044.

[86] 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.

[87] 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.

[88] Kellenberger, E., Foata, N., and Rognan, D. (2008) Ranking targets in structure-based virtual screening of three-dimensional protein libraries: methods and problems. *J Chem Inf Model 48*, 1014–1025.

[89] Gold, N. D. and Jackson, R. M. (2006) Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J Mol Biol 355*, 1112–24.

[90] Gold, N. D. and Jackson, R. M. (2006) SitesBase: a database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Res 34*, D231–4.

[91] Stark, A., Sunyaev, S., and Russell, R. B. (2003) A model for statistical significance of local similarities in structure. *J Mol Biol 326*, 1307–1316.

[92] Gold, N. D., Deville, K., and Jackson, R. M. (2007) New opportunities for protease ligand-binding site comparisons using SitesBase. *Biochem Soc Trans 35*, 561–565.

[93] Snyder, K. A., Feldman, H. J., Dumontier, M., Salama, J. J., and Hogue, C. W. V. (2006) Domain-based small molecule binding site annotation. *BMC Bioinformatics 7*, 152.

[94] Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., Buzadzija, K., Cavero, R., D'Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M. J., Dumontier, M. R., Earles, V., Farrall, R., Feldman, H., Garderman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa, A., Haw, R., Hrvojic, A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakis, J., Montojo, J., Moore, S., Muskat, B., Ng, I., Paraiso, J. P.,

Parker, B., Pintilie, G., Pirone, R., Salama, J. J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B. F. F., and Hogue, C. W. V. (2005) The Biomolecular Interaction Network Database and related tools 2005 update.'. *Nucleic Acids Res 33*, D418–D424.

[95] Michalsky, E., Dunkel, M., Goede, A., and Preissner, R. (2005) SuperLigands - a database of ligand structures derived from the Protein Data Bank. *BMC Bioinformatics 6*, 122.

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

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

[98] 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.

[99] 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.

[100] 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.

[101] 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.

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

[103] 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.

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

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

[106] 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.

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

[108] 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.

[109] 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.

[110] Stebbings, L. A. and Mizuguchi, K. (2004) Homstrad: recent developments of the homologous protein structure alignment database. *Nucleic Acids Res. 32*, D203–D207.

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

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

[113] 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.

[114] 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.

[115] 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.

[116] 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.

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

[118] 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.

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

[120] 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.

[121] 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.

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

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

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

[125] 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.

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

[127] (2007) Molecular operating environment (moe), 2007.08 (Chemical Computing Group, Montreal, C.N.).

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

[129] (2002) Sas, release 9.1 (SAS Institute Inc.: Cary, N.C.).

[130] (2007) Jmp, release 7.01 (SAS institute Inc.: Cary, N.C.).

[131] 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.

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

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

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

[135] 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.

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

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

[138] 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.

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

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

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

[142] 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.

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

[144] 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.

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

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

[147] 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.

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

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

[150] 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.

[151] 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.

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

[153] 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.

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

[155] 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.

[156] Echols, N., Milburn, D., and Gerstein, M. (2003) MolMovDB: analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Res 31*, 478–482.

[157] Freire, E. (1999) The propagation of binding interactions to remote sites in proteins: analysis of the binding of the monoclonal antibody D1.3 to lysozyme. *Proc Natl Acad Sci U S A 96*, 10118–22.

[158] Murray, C. W., Baxter, C. A., and Frenkel, A. D. (1999) The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *J Comput Aided Mol Des 13*, 547–562.

[159] Zhao, Y. and Sanner, M. F. (2008) Protein-ligand docking with multiple flexible side chains. *J Comput Aided Mol Des 22*, 673–679.

[160] May, A. and Zacharias, M. (2008) Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: evaluation on kinase inhibitor cross docking. *J Med Chem 51*, 3499–3506.

[161] Koska, J., Spassov, V. Z., Maynard, A. J., Yan, L., Austin, N., Flook, P. K., and Venkatachalam, C. M. (2008) Fully automated molecular mechanics based induced fit protein-ligand docking method. *J Chem Inf Model 48*, 1965–1973.

[162] Erickson, J. A., Jalaie, M., Robertson, D. H., Lewis, R. A., and Vieth, M. (2004) Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J Med Chem 47*, 45–55.

[163] 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.

[164] Gutteridge, A. and Thornton, J. (2005) Conformational changes observed in enzyme crystal structures upon substrate binding. *J Mol Biol 346*, 21–8.

[165] Brylinski, M. and Skolnick, J. (2008) What is the relationship between the global structures of apo and holo proteins? *Proteins 70*, 363–77.

[166] Schnecke, V. and Kuhn, L. (1999) *Database screening for HIV protease ligands: The influence of binding-site conformation and representation on ligand selectivity*. pp. 242–251.

[167] Schnecke, V. and Kuhn, L. (2000) Virtual screening with solvation and ligand-induced complementarity. *Perspectives in Drug Discovery and Design 20*, 171–190.

[168] 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.

[169] Benson, M. L., Smith, R. D., Khazanov, N. A., Dimcheff, B., Beaver, J., Dresslar, P., Nerothin, J., and Carlson, H. A. (2008) Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res 36*, D674–D678.

[170] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol 215*, 403–410.

[171] 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.

[172] 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.

[173] 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.

[174] Damm, K. L. and Carlson, H. A. (2007) Exploring experimental sources of multiple protein conformations in structure-based drug design. *J Am Chem Soc 129*, 8225–8235.

[175] Shoichet, B. K., McGovern, S. L., Wei, B., and Irwin, J. J. (2002) Lead discovery using molecular docking. *Curr Opin Chem Biol 6*, 439–446.

[176] Joseph-McCarthy, D., Baber, J. C., Feyfant, E., Thompson, D. C., and Humblet, C. (2007) Lead optimization via high-throughput molecular docking. *Curr Opin Drug Discov Devel 10*, 264–274.

[177] Rajamani, R. and Good, A. C. (2007) Ranking poses in structure-based lead discovery and optimization: current trends in scoring function development. *Curr Opin Drug Discov Devel 10*, 308–315.

[178] Seifert, M. H. J., Kraus, J., and Kramer, B. (2007) Virtual high-throughput screening of molecular databases. *Curr Opin Drug Discov Devel 10*, 298–307.

[179] Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005) The amber biomolecular simulation programs. *J Comput Chem 26*, 1668–1688.

[180] Wang, W., Donini, O., Reyes, C. M., and Kollman, P. A. (2001) Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annu Rev Biophys Biomol Struct 30*, 211–243.

[181] Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A., and Honig, B. (2002) Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J Comput Chem 23*, 128–137.

[182] Grant, J. A., Pickup, B. T., and Nicholls, A. (2001) A smooth permittivity function for poisson-boltzmann solvation methods. *J. Comput. Chem. 22*, 608–640.

[183] Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A 98*, 10037–10041.

[184] Wei, B. Q., Baase, W. A., Weaver, L. H., Matthews, B. W., and Shoichet, B. K. (2002) A model binding site for testing scoring functions in molecular docking. *J Mol Biol 322*, 339–355.

[185] Bashford, D. and Case, D. A. (2000) Generalized born models of macromolecular solvation effects. *Annu Rev Phys Chem 51*, 129–152.

[186] Chen, J., Brooks, C. L., and Khandogin, J. (2008) Recent advances in implicit solvent-based methods for biomolecular simulations. *Curr Opin Struct Biol 18*, 140–148.

[187] Zou, X., Sun, Y., and Kuntz, I. (1999) Inclusion of solvation in ligand binding free energy calculations using the generalized-born model. *J. Am. Chem. Soc. 121*, 8033–8043.

[188] Liu, H., Kuntz, I., and Zou, X. (2004) Pairwise GB/SA scoring function for structure-based drug design. *J. Phys. Chem. B. 108*, 5453–5462.

[189] Liu, H.-Y. and Zou, X. (2006) Electrostatics of ligand binding: parametrization of the generalized Born model and comparison with the Poisson-Boltzmann approach. *J Phys Chem B 110*, 9304–9313.

[190] Jain, A. N. (1996) Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. *J Comput Aided Mol Des 10*, 427–440.

[191] Head, R., Smythe, M., Oprea, T., Waller, C., Green, S., and Marshall, G. (1996) Validate: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc. 118*, 3959–3969.

[192] Gehlhaar, D. K., Verkhivker, G. M., Rejto, P. A., Sherman, C. J., Fogel, D. B., Fogel, L. J., and Freer, S. T. (1995) Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem Biol 2*, 317–324.

[193] Gehlhaar, D., Bouzida, D., and Rejto, P. (1999) *In Rational Drug Design: Novel Methodology and Practical Applications*.

[194] Tanaka, S. and Scheraga, H. A. (1976) Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules 9*, 945–950.

[195] Miyazawa, S. and Jernigan, R. (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules 18*, 534–552.

[196] Sippl, M. J., Lackner, P., Domingues, F. S., and Koppensteiner, W. A. (1999) An attempt to analyse progress in fold recognition from CASP1 to CASP3. *Proteins Suppl 3*, 226–30.

[197] Vajda, S., Sippl, M., and Novotny, J. (1997) Empirical potentials and functions for protein folding and binding. *Curr Opin Struct Biol 7*, 222–228.

[198] Verkhivker, G., Appelt, K., Freer, S. T., and Villafranca, J. E. (1995) Empirical free energy calculations of ligand-protein crystallographic complexes. i. knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng 8*, 677–691.

[199] Mitchell, J., Laskowski, R., Alex, A., and Thornton, J. (1999) BLEEP - potential of mean force describing protein-ligand interactions: I. generating potential. *Journal of Computational Chemistry 20*, 1165–1176.

[200] Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. J., and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A 97*, 262–7.

[201] Ishchenko, A. V. and Shakhnovich, E. I. (2002) SMall Molecule Growth 2001 (SMoG2001): an improved knowledge-based scoring function for protein-ligand interactions. *J Med Chem 45*, 2770–2780.

[202] Zhang, C., Liu, S., Zhu, Q., and Zhou, Y. (2005) A knowledge-based energy function for protein-ligand, protein-protein, and protein-dna complexes. *J Med Chem 48*, 2325–2335.

[203] Muegge, I. (2006) Pmf scoring revisited. *J Med Chem 49*, 5895–5902.

[204] Huey, R., Morris, G. M., Olson, A. J., and Goodsell, D. S. (2007) A semiempirical free energy force field with charge-based desolvation. *J Comput Chem 28*, 1145–1152.

[205] Huang, S.-Y. and Zou, X. (2007) Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins 66*, 399–421.

[206] Huang, S.-Y. and Zou, X. (2007) Efficient molecular docking of NMR structures: application to HIV-1 protease. *Protein Sci 16*, 43–51.

[207] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem 25*, 1605–1612.

[208] Gasteiger, J. and Marsili, M. (1980) Iterative partial equalization of orbital electronegativitya rapid access to atomic charges. *Tetrahedron 36*, 3219–3228.

[209] Shi, D., Morizono, H., Ha, Y., Aoyagi, M., Tuchman, M., and Allewell, N. M. (1998) 1.85-a resolution crystal structure of human ornithine transcarbamoylase complexed with n-phosphonacetyl-l-ornithine. catalytic mechanism and correlation with inherited deficiency. *J Biol Chem 273*, 34247–34254.

[210] Sugiyama, S., Matsuo, Y., Maenaka, K., Vassylyev, D. G., Matsushima, M., Kashiwagi, K., Igarashi, K., and Morikawa, K. (1996) The 1.8-a x-ray structure of the escherichia coli potd protein complexed with spermidine and the mechanism of polyamine binding. *Protein Sci 5*, 1984–1990.

[211] Zanotti, G., Marcello, M., Malpeli, G., Folli, C., Sartori, G., and Berni, R. (1994) Crystallographic studies on complexes between retinoids and plasma retinol-binding protein. *J Biol Chem 269*, 29613–29620.

[212] Ferrer, J. L., Jez, J. M., Bowman, M. E., Dixon, R. A., and Noel, J. P. (1999) Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis. *Nat Struct Biol 6*, 775–784.

[213] Campanacci, V., Bishop, R. E., Blangy, S., Tegoni, M., and Cambillau, C. (2006) The membrane bound bacterial lipocalin blc is a functional dimer with binding preference for lysophospholipids. *FEBS Lett 580*, 4877–4883.

[214] Angelucci, F., Johnson, K. A., Baiocco, P., Miele, A. E., Brunori, M., Valle, C., Vigorosi, F., Troiani, A. R., Liberti, P., Cioli, D., Klinkert, M.-Q., and Bellelli, A. (2004) Schistosoma mansoni fatty acid binding protein: specificity and functional control as revealed by crystallographic structure. *Biochemistry 43*, 13000–13011.

[215] Price, A. C., Choi, K. H., Heath, R. J., Li, Z., White, S. W., and Rock, C. O. (2001) Inhibition of beta-ketoacyl-acyl carrier protein synthases by thiolactomycin and cerulenin. structure and mechanism. *J Biol Chem 276*, 6551–6559.

[216] Lee, J. O., Yang, H., Georgescu, M. M., Cristofano, A. D., Maehama, T., Shi, Y., Dixon, J. E., Pandolfi, P., and Pavletich, N. P. (1999) Crystal structure of the pten tumor suppressor: implications for its phosphoinositide phosphatase activity and membrane association. *Cell 99*, 323–334.

[217] Johnston, J. M., Arcus, V. L., and Baker, E. N. (2005) Structure of naphthoate synthase (menb) from mycobacterium tuberculosis in both native and product-bound forms. *Acta Crystallogr D Biol Crystallogr 61*, 1199–1206.

[218] Falke, J. J., Dernburg, A. F., Sternberg, D. A., Zalkin, N., Milligan, D. L., and Koshland, D. E. (1988) Structure of a bacterial sensory receptor. a site-directed sulfhydryl study. *J Biol Chem 263*, 14850–14858.

[219] Rojas, J. R., Trievel, R. C., Zhou, J., Mo, Y., Li, X., Berger, S. L., Allis, C. D., and Marmorstein, R. (1999) Structure of tetrahymena gcn5 bound to coenzyme a and a histone h3 peptide. *Nature 401*, 93–98.

[220] Hoff, K. G. and Wolberger, C. (2005) Getting a grip on O-acetyl-ADP-ribose. *Nat Struct Mol Biol 12*, 560–561.

[221] Lesniak, J., Barton, W. A., and Nikolov, D. B. (2002) Structural and functional characterization of the pseudomonas hydroperoxide resistance protein ohr. *EMBO J 21*, 6649–6659.

[222] Cho, H. S., Choi, G., Choi, K. Y., and Oh, B. H. (1998) Crystal structure and enzyme mechanism of delta 5-3-ketosteroid isomerase from pseudomonas testosteroni. *Biochemistry 37*, 8325–8330.

[223] Holland, D. R., Clancy, L. L., Muchmore, S. W., Ryde, T. J., Einspahr, H. M., Finzel, B. C., Heinrikson, R. L., and Watenpaugh, K. D. (1990) The crystal structure of a lysine 49 phospholipase a2 from the venom of the cottonmouth snake at 2.0-a resolution. *J Biol Chem 265*, 17649–17656.

[224] Siebold, C., Arnold, I., Garcia-Alles, L. F., Baumann, U., and Erni, B. (2003) Crystal structure of the citrobacter freundii dihydroxyacetone kinase reveals an eight-stranded alpha-helical barrel ATP-binding domain. *J Biol Chem 278*, 48236–48244.

[225] (.) http://www.jboss.org/products/jbossas.

[226] (.) http://java.sun.com/products/javabeans/.

[227] (.) http://struts.apache.org/.

[228] Gamma, E., Helm, R., Johnson, R., and Vlissides, J. (1995) *Design Patterns*. (Addison-Wesley Professional).

[229] Yamaguchi, A., Iida, K., Matsui, N., Tomoda, S., Yura, K., and Go, M. (2004) Het-PDB Navi.: a database for protein-small molecule interactions. *J.Biochem.(Tokyo, Jpn.) 135*, 79–84.

[230] 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.

[231] Andrade, M. A. and Bork, P. (2000) Automated extraction of information in molecular biology. *FEBS Lett 476*, 12–17.

[232] Shah, P. K., Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2003) Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics 4*, 20.

[233] Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. (2001) Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics 17*, 155–161.

[234] Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005) Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics 6 Suppl 1*, S1.

[235] Jurafsky, D., Martin, J., Kehler, A., Vander Linden, K., and Ward, N. (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. (MIT Press).