Structural Investigation of Binding Events in Proteins

by

Jordan J. Clark

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Medicinal Chemistry)
in the University of Michigan
2018

Doctoral Committee:

Professor Heather A. Carlson, Chair
Professor Charles L. Brooks III
Professor Henry I. Mosberg
Professor Nouri Neamati
Professor Janet L. Smith

Jordan J. Clark
Jordanjc@umich.edu
ORCID iD: 0000-0003-3863-4183

Table of Contents

List of Figures

List of Tables

List of Appendices

Abstract

Understanding the biophysical properties that describe protein binding events has allowed for the advancement of drug discovery through structure-based drug design and *in silico* methodology. The accuracy of these *in silico* methods depends entirely on the parameters that we determine for them. Many of these parameters are derived from the structural information we have obtained as a community and therein resides the importance of integrity of the quality of this structural data.

First, the curation and contents of the Binding MOAD database are extensively described. This database serves as a repository of 25,759 high-quality, ligand-bound X-ray protein crystal structures complemented by 9138 hand-curated binding affinity data for as many of those ligands as appropriate. The newly implemented extended binding site feature is presented, establishing more robust definitions of ligand binding sites than those provided by other databases. Finally, the contents of Binding MOAD are compared to similar databases, establishing the value of our dataset and which purposes it best serves.

Second, a robust dataset of 305 unique protein sequences with at least two ligand-bound and two ligand-free structures for each unique protein is cultivated from Binding MOAD and the PDB. Protein flexibility is assessed using C_α RMSD for backbone motion and χ_1 angles to quantify side-chain motions. We establish that there is no statistically significant difference between the available conformational space for the backbones or the side chains of unbound proteins when compared to their bound structures. Examining the change in occupied conformational space upon ligand binding reveals a statistically significant increase in backbone conformational space of miniscule magnitude, but a significant increase of side-chain conformational space. To quantify the conformational space available to the side chains, flexibility profiles are established for each amino acid. We found no correlation between backbone and side-chain flexibility. Parallels are then made to common practices in flexible docking techniques.

Six binding-site prediction algorithms are then benchmarked on a derivation of the previously established dataset of 305 proteins. We assessed the performance of ligand-bound vs ligand-free structures with these methods and concluded that five of the six methods showed no preference for either structure type. The remaining method, Fpocket, showed decreased performance for ligand-free structures. There was a staggering amount of inconsistency in performance with the methods; different structures of the exact same protein could achieve wildly different rates of success with the same method. The performance of individual structures for all six methods indicated that success and failure rates were seemingly random. Finally, we establish no correlation between the performance of the same structures with different methods, or the performance of the structures with structure resolution, Cruickshank DPI, or number of unresolved residues in their binding sites.

Last, we examine the chemical and physical properties of protein-protein interactions (PPIs) with regard to their geometric location in the interface. First, we found that the relative elevation changes of the protein interface landscapes demonstrate that these interfaces are not as flat as previously described. Second, the hollows of druggable PPI interfaces are more sharply shaped and nonpolar in nature, and the protrusions of these druggable PPI interfaces are very polar in character. Last, no correlations exist between the binding affinity describing the subunits of a PPI and other physical and chemical parameters that we measured.

Chapter 1. Introduction

Proteins are naturally flexible biopolymers capable of performing a wide variety of biochemical functions including signaling, protein processing, regulation, and alteration of small molecules (metabolism, for example). Many of these processes involve small molecules in the form of substrates, cofactors, and allosteric regulators. These binding events are critical for the field of medicinal chemistry because they present numerous targetable interactions amenable for engineering therapeutic agents. Due to the significant variety of biochemical and physical properties of both protein binding sites and the ligands that bind to them, there has been significant amounts of research and debate about the most influential factors in creating high-affinity ligands. These factors can be largely attributed either to the ligand, or to the protein.

This dissertation studies the binding interactions between proteins and their ligand targets by utilizing a large database of high-quality, X-ray crystal structures of protein-ligand complexes called Binding MOAD.[1-3] This study is completed in a very protein-centric manner, concentrated on the different aspects of protein flexibility and how the shape of proteins is influenced by ligand binding events. The bulk of the investigation is preceded by introduction of new, more robust definitions of binding sites in Binding MOAD. The main study of protein flexibility is then accomplished, and the dataset is then further utilized in a brief survey of ligand binding site prediction algorithms to test how well they can replicate these more robustly defined binding sites. Lastly, this dissertation presents a basic study of protein-protein interface topography.

1.1 Protein Flexibility

Proteins are naturally flexible biomacromolecules made up of amino acids. This flexibility comes as a result of the nearly infinite combinations and arrangements of the amino acids it is composed of. Though proteins are polymeric in assembly, they primarily exist in folded conformations guided and stabilized by non-covalent interactions. Their flexibility is a vital

component of binding substrates, performing enzymatic catalysis, and releasing bound substrates.[4-5]

Studying the flexibility of proteins and their binding sites has yielded some key insights of what governs protein-ligand interactions (PLIs). Heringa and Argos' study of PLIs yielded some cases of binding induced strain in clusters of residues in or near ligand binding sites.[6] This strain manifested as non-rotameric side-chain strain, meaning the side chains were occupying high energy conformations, but not actually switching to the next rotameric position. They hypothesized that side chains occupying strained conformations away from the energetic minimum allowed may help to drive enzymatic reactions.[7] Luque and Freire describe how protein binding sites are often characterized by regions of both high and low stability.[8-9] Some of the instable regions have been shown to be absolutely necessary for proper protein function.[10] Catalytic residues in enzymes are typically located in highly stable regions of the protein core, which may allow for some preorganization of the binding site. Low stability regions have been shown to play a role in communication between allosteric binding sites and primary active sites. This preorganization concept has been investigated heavily, and its extent is subject to heavy debate within the community.[11]

1.2 Protein-Ligand Binding

Theories about small molecule interactions with proteins have changed constantly over the past century and continue to evolve. In 1894, the "lock and key" model was proposed by Herman Emil Fisher, describing the shape of ligand binding pockets as being predetermined in nature and only allowing ligands of the proper complementary shape to bind.[12-14] In 1948, Linus Pauling suggested that enzymes were complementary molecules in structure to the activated complexes for the reactions which they catalyze. [15-16] These ideas were built upon and the "induced fit' model was proposed in 1958 by Daniel Koshland, suggesting that protein binding sites adjust to accommodate ligands.[17] Current theories range across that gamut, but many acknowledge that proteins exist in an equilibrium of energetically similar states.[4, 9, 18-20] Ligands may then bind and "trap" proteins in a desired conformational state, which may shift the equilibrium of the system towards a distribution of conformations more favorable to the binding reaction.[21]

Ligand binding events are often quantified using the free energy of binding for a ligand ($\Delta G_{binding}$). This free energy of binding, given by the relationship: $\Delta G_{binding} = \Delta H_{binding} - T\Delta S_{binding} = -RT\ \ln(K_a)$, where enthalpy ($\Delta H$) and entropy ($\Delta S$) are the state variables that govern the interaction, which is often measured by the equilibrium constant between the protein and ligand (K_a). The specific contribution of entropy and enthalpy is dependent on the identity of the protein and ligand, as well as the conditions (i.e. concentrations of other molecules and solvents present). Differing solvent conditions for separate crystal structures only exacerbate this problem, as solvent has been shown to have a massive impact on various dynamics of protein structures.[22] Several factors govern the determination of both the entropy and enthalpy of a protein-ligand binding event, but these values are inherently difficult to calculate accurately.[13, 23] Due to this, comparison between binding constants of different types (IC_{50}, K_i, K_d) can yield wildly varying results.

1.2.1 Solvation and Desolvation

Biological processes such as ligand binding occur in aqueous environments, and that water plays a significant role in the binding process. As it is the bulk solvent, water surrounds basically every component of a protein as well as free ligands at equilibrium. Therefore, that water must be displaced from both the protein's binding site, as well as the ligand's surface before ligand binding is possible.[13] Waters inside of protein binding sites are typically partially occupied and able to associate/dissociate freely. Desolvation can be both favorable and unfavorable to binding, and usually correlates to the polarity of the entity being desolvated. Desolvating charged groups is almost always unfavorable to binding[24], while the hydrophobic effect results in desolvation as a favorable process to binding.

The hydrophobic effect was first discussed in 1945 by Frank and Evans, as a means to explaining the positive influence non-polar molecules can have on the free energy of binding, despite the desolvation that must occur.[25] Introduction of a non-polar molecule to an aqueous environment is normally an energetically unfavorable process.[26] This is due to disruption of the network of hydrogen bonds between water molecules, around where the non-polar molecule is located. However, in the case of a protein in an aqueous environment, a non-polar molecule can bury itself within the protein and the external water is able to rearrange back to a favorable

hydrogen-bonding network, resulting in a positive impact on the free energy of binding.[27] It has been shown that solvent reorganization can attribute anywhere from 25% to 100% of the enthalpy gained in small-molecule binding.[28] It has also been shown that the enthalpic contribution of the hydrophobic effect is proportional to the amount of buried non-polar surface area.[27]

1.2.2 Van der Waals Interactions and Electrostatic Interactions

Van der Waals (VdW) interactions are one of the most significant contributors to high-affinity binding events with proteins.[29] These low-energy interactions are created by London Dispersion forces when placing atoms in contact with each other. Binding sites tend to be buried cavities, with little solvent exposure.[30] This idea of buried-ness becomes important in high-affinity ligand binding events to ensure the maximum amount of surface contact with the ligand. The "lock and key" ligand binding model is based on this principle, as a perfect fit is only attainable when one specific ligand is involved.

Electrostatic interactions are a broad classification of polarized molecular interactions which include hydrogen bonds, salt bridges, and metal contacts. Electrostatic interactions are less common than VdW interactions but have a much larger impact on the enthalpy of binding. Strong hydrogen bonds form between a polar-atom-bound hydrogen and another polar atom, where the polar atoms are commonly an O or an N. These stronger hydrogen bonds generally contain an enthalpic contribution 3-7 kcal/mol, implying the amount of energy necessary to break the bond. Weaker hydrogen bonds also exist, where the hydrogen-bound atom is not of electronegative or polar origin. An example of this weaker hydrogen bonding is a CH-O hydrogen bond, which typically contain only 1-2 kcal/mol, outside of examples with charged species.[31] These weaker hydrogen have been crystallographically observed[32] and also experimentally observed by means of solvent boiling temperature comparisons of halogenated and hydrogenated solvents.[16] Importantly, the strength of any hydrogen bond is highly dependent on its geometry.[13, 33]

In a biological context, hydrogen bonding with water must be considered when describing ligand binding events. Both ligands and proteins are constantly interacting with water, as it is the bulk solvent in most biological environments. In order for PLIs to occur, both the protein and the ligand must undergo desolvation, which has both an enthalpic and entropic cost. Since the free energy of binding represents total difference between the standard free energy of the ligand-free protein and the standard free energy of the ligand-bound complex, and the chemical space of the

entire protein is typically much larger than the chemical space of the binding site, it is believed that the specifically observable PLIs do not greatly contribute to the free energy of binding.[13] Past work has shown that expanding a ligand's footprint to achieve another hydrogen bond may not yield any improvement in binding affinity.[34] This occurred because the enthalpic gain of an additional hydrogen bond can be outweighed by the entropic cost of desolvation of the protein's polar group participating in the hydrogen bond, as well as the entropic cost of forcing a specific orientation for that polar group.

Lastly, salt bridges are the strongest class of electrostatic interaction. Salt bridges are formed between a positively charged and negatively charged set of functional groups. Despite the strength of salt bridges, the desolvation penalty of removing water from a charged group is still quite large.[35]

Despite knowing the individual strengths of these different types of interactions, their contribution to the free energy of binding in real practice is far more convoluted. There are many arguments for which interaction types contribute the most towards extremely tight binding events. An important set of studies focused on the binding of biotin to streptavidin, as it is the tightest known protein-ligand complex discovered at the time and still is now, 25 years later. Early calculations based on free energy perturbation indicated that the extreme binding affinity in the biotin-streptavidin system was due to van der Waals contacts and also suggested that the binding pocket for biotin was preformed, supporting the lock and key theory of binding.[36] A newer study combined quantum/molecular mechanics and Monte Carlo computational techniques on hydrogen-bonding residues in streptavidin, and revealed that networks of hydrogen bonding were responsible for the strong binding of the biotin-streptavidin complex.[37] Later work confirmed the presence of a sophisticated hydrogen bonding network using isothermal calorimetry, revealing an 11-fold greater contribution to the free energy of binding for two coupled residues involved in hydrogen bonding when comparing to what their individual contributions would be, otherwise.[38]

1.2.3 Defining Protein-Ligand Binding Sites and Unified Binding Sites

In most studies based on structural biology, protein binding sites are derived from protein-ligand contacts using the bound ligands in the crystal structures used. This approach has the benefit of only providing contacts which are physically observed in the crystal structure, which is itself a set of data derived from an actual biophysical experiment. Some may deem these to be "real"

contacts where other suspected contacts are questionable. While this benefit is appreciable, the consequential assumption of relevant binding residues is heavily influenced by the number of ligand-bound structures contained in the dataset used for a given experiment.

Studies using smaller, younger datasets sometimes dealt with the lack of sufficient data to represent either ligand binding state of a protein (apo or holo) by using apo-holo protein pairs as a representation of observable conformational change upon ligand binding. While this approach may help reveal appreciable differences in protein conformation tied to its ligand binding state, it doesn't reveal inherent flexibility of either ligand binding state. Incorporating multiple of each structure type is therefore advantageous to describing the inherent variance present both with and without ligands bound.

When incorporating multiple protein structures for binding site definition, the binding site contacts for multiple different ligands in different structures must be combined in some manner, and two approaches are immediately apparent: an exclusive method, and an inclusive method. With exclusive contact merging, only residues commonly contacted by all included ligands would be considered. This more restrictive definition could be useful in identifying necessary binding residues in a very rigid context, such as an enzymatic catalytic site where only crucial residues immediately surrounding a catalytic ensemble are to be identified. However, when attempting to describe biophysical properties of an 'entire' binding site, this method is undesirable. The converse approach is an 'inclusive' method, where all of the binding site residues contact by any ligand of the represented protein are considered as part of the binding site.

This more complete definition of a binding site can be viewed as a 'union' of all representations of a binding site for a given protein sequence, and we will therefore refer to it as a Union Binding Site (UBS). The UBS concept is heavily utilized in Chapters 2-3 of this dissertation, and some of its implications are probed in Chapter 4.

1.2.4 **Predicting Protein-Ligand Binding Sites**

Another useful application of robust binding site knowledge is in predicting binding site locations where they may not be known. With the increasing popularity of structure-based drug design, acquisition of structural information for relevant new targets has become even more important. If the structural information is obtained early in the investigation of a new target, the

relevant binding cavities may not be known, especially in cases where there are allosteric binding sites present.

Ligand binding-site prediction methods are divided into four categories for discussion: template-based methods (sometimes referred to as genomic-based methods), geometry-based methods, energy-based methods, and other methods.

Template-based methods utilize the atlas of already known protein information as a protein roadmap to guide the detection algorithm. Their assumption is that binding sites of new protein sequences may be located using the known binding sites of close structural homologs.

Geometry-based methods explore and characterize protein surfaces using a number of biophysical parameters such as Van der Waals radii to locate pockets or clefts. Most geometric methods assume that the binding site of a protein is a cleft or pocket in the protein surface. Exploration of the protein surface may be accomplished by calculation of molecular distance, solvent accessible surface area (SASA), and cavity volume. These measurements are computed using probes, spheres, grids, and other forms of spatial voids, which are then clustered or further analyzed to yield ranked cavities presumed to be binding sites.

Energy-based methods rely on calculation of phenomena such as hydrogen bonding and pi-stacking to locate regions of the protein where ligands are likely to bind. These LBS-prediction methods are expected to be relatively quick in terms of computation time, so energy-based methods must only account for simple phenomena or make many assumptions to reduce the number of calculations necessary. These methods utilize probe molecules and chemical moieties to generate potentials which are then scored to locate binding sites.

Methods may be categorized as "other" methods, because they either use a different set of physicochemical phenomena not discussed previously, or are a combination of different approaches.[39] Examples of new parameters include sequence-conservation, use of machine learning protocols, targeted-sequence (i.e. finding residues that bind metal ions), ligand-centric methods which use some derivation of a pharmacophore model to search for binding sites, and 'meta-analyses.' Meta analyses combine any number of the previous types of methodology or even full methods and use a scoring algorithm to combine the results of the different protocols and rank them to find the best predicted site that multiple methods can agree on.

While there are many types of binding-site prediction algorithms, their rates of success can vary wildly over different targets. Robust methodology such as molecular dynamics simulations

using small molecule fragments are unquestionably better at predicting relevant binding locations of protein surfaces, but binding-site prediction methods are expected to be fast and accessible. Individual binding-site prediction methods and further expansion of each method type can be found in section 4.2.2.

1.3 Protein-Protein Binding

Protein-protein binding events, often discussed as protein-protein interactions (PPIs), occur between two or more protein chains. These chains may be identical or non-identical in sequence, and the duration of interaction may be either transient or permanent, as well as obligate or non-obligate in nature. Furthermore, transient PPIs can be either strong or weak, which is believed to correlate strongly with the stand-alone stability of the individual monomers.[40] Due to the large number of potential types of interactions and the many dynamic physical features of protein interfaces, PPIs are some of the most complicated macromolecular interactions known to biochemists.[40-41]

PPIs can be arrange in both an isologous or heterologous way with respect to structural symmetry.[40] Isologous association involves the same surface on both monomers, which relies on them either being identical chains, or closely related structural homologs. Heterologous association involves binding to a different interfacial position for each monomer, which results in infinite aggregation outside of an arrangement with cyclic symmetry.

The interaction between the protein chains can be broken down into chemical and geometric analysis.[42] Large scale geometry of PPIs is primarily an analysis of structural orientation. Interactions of dimeric protein pairs vary from simple, flat interfaces with an oval shaped contact patch, to situations where the larger of the pair completely engulfs the smaller protein or peptide chain. Finer geometric aspects of the interaction describe the intertwined or interdigitated nature of the complexed protein-protein chains. Surface complementarity[43] has been utilized in attempt to quantitatively describe the intertwined nature of the interface, but the topographical nature of it lacks qualitative definition when using such a broad metric. While interfaces are generally deemed as symmetrical and therefore "flat and featureless" [41, 44-46], PPIs do contain topographically interesting features of "protrusions" that pass across the theoretical center plane of an interface, burying themselves in a "hollow" in the adjacent protein chain.[43, 47]

Chemical properties of PPIs are relatively similar to the analysis of protein-ligand interaction sites. These analyses consist of calculating the relative frequencies of the various types of amino acids or atoms: polar, nonpolar, and charged.

While obvious differences between protein-protein binding hot spots and protein-ligand binding exist, there are many parallels. The hotspots for binding proteins also rely on rigid and flexible residues to bind and eject binding partners. Nussinov showed that rigid residues may be used as anchors, while surrounded by flexible residues. [48-50]

1.3.1 Defining PPI Interfaces

The issue of accurately defining the residue contacts of a PPI stems from both the lack of structural data of the complexed pair of proteins and the characteristics of protein-protein binding events. The nature of binding between protein partners is far less promiscuous than PLIs, as most protein-binding proteins only have one desired partner in a biological context. However, due to the either extremely stable, or instable nature of PPIs, structural information for either the complexed or uncomplexed state of the binding partners is often elusive. Furthermore, the affinity of a protein complex can be altered by changes in the concentration of ions and molecules in solution, pH, temperature, covalent modification and, of course, the presence of molecules capable of binding to either protein in the complex.[40] These details further complicate the acquisition of structural information for PPIs, since they are the very techniques that crystallographers use to create appropriate buffer solutions to stabilize protein crystal formation.

The characteristics of protein-protein binding are also different than protein-ligand binding in numerous ways, first and foremost by the sheer size of the interaction area. Due to this large interaction area, the most prevalent method of determining interaction area has been to track the loss of solvent accessible surface area (SASA) from residues upon complexation.[43, 51-54] The features of this methodology and their consequences will be discussed further in Chapter 5.

1.4 Use of databases

Due to the inherent variation of proteins and their adapted conformations in X-ray crystal structures, it is important to investigate a wide range of proteins bound to a variety of ligands. Different proteins having different contributions to free energy of binding is unsurprising, and

necessary. Databases of these interactions are a necessary component to completing our research, because we aim to make generalized statements regarding a wide variety of diverse protein-ligand (Chapters 2-4), and protein-protein (Chapter 5) complexes. Structural coordinates are vital to reinforcing our claims, so databases focusing on structural content are of highest importance. Pairing affinity information with the structural information provides an invaluable opportunity to test the correlative nature of binding strength against many physicochemical properties of the corresponding protein targets. This work focuses primarily on the use of Binding MOAD as a central resource of protein-ligand crystal structures. A few other databases, namely BindingDB, PDBbind, and AffinDB contain similar information. The deficiencies of these databases are discussed in depth in Chapter 2, as well addressing a few other related databases. The two databases used heavily in dataset creation for the experimental work in this dissertation are briefly introduced below.

1.4.1 **Binding MOAD**

Binding MOAD is a collection of high quality, X-ray crystal structures of protein-ligand complexes maintained by the Carlson laboratory at the University of Michigan. This database aims to couple binding data with structural information to assist with drug-discovery centric research and the study of PLIs in a broad biophysical context. The construction, maintenance, and contents of this database are discussed in depth during Chapter 2. A few other databases with distinct protein-ligand information are discussed more in depth during Chapters 2 and 3.

1.4.2 **2P2I DB**

The protein-protein interaction inhibitors database (2P2I db)[55-56] is dedicated to hand-curated structural information of PPIs with known inhibitors. This database is curated primarily from the PDB and focuses on inhibition of heteromeric protein-protein complexes. The dataset aims to contain the complexed-state, and a ligand-bound structure for at least one of the individual protein partners, or a very close homologue if no direct matches are available. The original release of 2P2I db consisted of 17 protein-protein complexes representing 14 families and 56 small molecule inhibitors.[56] The database is updated regularly, far more frequently than there are corresponding publications. The most recent update contains 31 PPIs and 242 small molecule

inhibitors.[57] Curation and other details of 2P2I db is discussed extensively in section 5.2.4.1. A few other databases with distinct PPI information are presented in Chapter 5.

1.5 Overview of thesis

The major areas addressed in this dissertation include expansion of binding site data contained in Binding MOAD, probing of ligand-binding induced flexibility in proteins and their binding sites, survey of geometrically based binding-site prediction algorithms, and rudimentary mapping of protein-interface topography.

Chapter 2 describes Binding MOAD, including data acquisition and updating procedures. The major contribution of this body of work to Binding MOAD is the introduction of unified binding-sites to the database. This function is utilized heavily in Chapters 3 and 4.

Chapter 3 presents a large-scale study on ligand-binding induced flexibility of protein backbones, as well as side-chain flexibility of their binding sites. This is the first study of its kind to probe the inherent flexibility of both control groups: ligand-bound structures, and ligand-free structures, as well as contrasting between the two states. The dataset includes 4048 protein structures, 2369 ligand-bound (holo), and 1679 ligand-free (apo), collectively divided into 305 protein families which each contain at least 2 apo and 2 holo structures.

Chapter 4 presents a survey of six binding-site prediction algorithms against a condensed version of the dataset created in Chapter 3. The intent of this research is to probe the performance differences between ligand-bound and ligand-free crystal structures with ligand binding-site prediction methodology. Relationships between structure quality and performance are also investigated.

Finally, Chapter 5 introduces a geometric method of analysis for protein-protein interfaces, to describe their localized surface topography as an attempt to better understand their unique binding characteristics. Both chemical and physical properties are attributed in the geometric space of the interface surface, and correlations between various characteristics are investigated.

Chapter 2. Binding MOAD (Mother of All Databases)

2.1 Introduction

Studies across biochemical disciplines regularly utilize datasets of macromolecular structure for their targets of interest. Often, these data are X-ray crystal structures of protein targets acquired from the Protein Data Bank (PDB).[58-59] Early protein datasets were small enough to exist only as a list of relevant PDB IDs inside of their publication. As the amount of data utilized in these types of studies has increased from mere tens of structures to the hundreds or even thousands of structures employed in more modern publications, the list sizes are too large to be included in the main body-text. This has resulted in datasets presented as separate downloadable entities or even hosted on the web as publicly accessible tools. Publicly available resources are of unquestionable use to the scientific community, so long as they are maintained regularly and transparently described in their or original publication as to be reproducible and accurately utilized.

Binding MOAD is a database of carefully curated, high quality, protein-ligand crystal structures of biologically interesting small molecules. This database includes binding data for many of the ligand-protein pairs, curated from their primary citation. The database is accessible *via* the web at www.BindingMOAD.org. Data is presented to users on a per-structure basis, but the proteins are also grouped by various sequence identity cutoffs to make finding similar structures easier. A few different versions of the dataset are available for download on the downloads page. This includes a version with only the curated binding data, as well as a fully compressed and zipped copy of the collective biological unit files for all entries.

Our aim is to make Binding MOAD the largest resource of high-quality, protein-ligand complexes available from the Protein Data Bank and augment that set with appropriate binding data as well as tools for finding similar binding sites and binding partners. When initially introduced in 2005, Binding MOAD contained 5331 protein-ligand complexes, augmented with 1375 binding data for 26% of the protein-ligand complexes.[2] Currently, Binding MOAD contains 25,759 protein-ligand structures with 12,432 different ligands, for which we have 9,142 binding data. The focus of this chapter is the introduction of unified binding sites to incorporate data

redundancy in a meaningful way to Binding MOAD. These uniquely robust binding sites have great potential to fortify *in silico* methodology, providing additional data for binding site prediction algorithms. Binding MOAD will be the first database to carry this variety of extended binding site information. A few examples of the most valuable useful and largest related databases are outlined below. Strengths and weaknesses of each database are noted, and the comparative utility of Binding MOAD is highlighted.

2.2 PDBbind

PDBbind was originally created by Shaomeng Wang and coworkers and is now maintained by Renxiao Wang and coworkers.[60-61] It contains 17,900 total complexes with binding data, 14,761 of those consisting of protein-small molecule ligand pairings, which is referred to as the "general set." This general set is then reduced to a "refined set" utilizing a number of cutoffs: only X-ray crystal structures, requiring a resolution better than 2.5 Å, R-factor lower than 0.250, all fragments of ligand molecules must be present, all backbone and sidechain fragments of the protein binding sites (defined within 8 Å of the ligand) must be present, binding constants must be between 1 pM and 10 mM and must be absolute measurements (i.e. $K_d \sim 1$ nM is not accepted), with ligands <1000 g/mol molecular weight, peptides ≤ 10 amino acids, polynucleotides < 4 residues, and ligand buried surface area must be > 15% of the total ligand surface area.[62] This refined set contains 4,154 protein-ligand complexes.

The refined set is then further reduced to a "core set," which aims to represent each present protein sequence with three members, requiring a 100-fold difference in binding affinity between the three structures (10 fold between each structure pair). The workflow for this dataset follows:

1. The refined set is binned by 90% sequence identity and only families with >4 members are kept
2. The binding constants within the families are compared, requiring a 10-fold difference between the minimum and median, and 10-fold difference between the median and maximum (maximum being the tightest binder)
3. The electron density fit a number of criteria, including complete and proper envelopment of the ligand as well as a lack of "too much" positive and/or negative electron density within the ligand binding site.

a. Systems which have a maximum/median/minimum structure fail step 3, are then replaced with the next closest family member in the category from step 2, when available, in attempt to keep as much data as possible.

This final core set contains 285 complexes which represents 95 unique proteins, as each represented protein has three structures.

PDBbind's recent efforts have been focused on expanding their most rigorous dataset, the core set, to increase the amount of available data to be used in their scoring function benchmark, the Comparative Assessment of Scoring Functions (CASF).[63] Binding MOAD's somewhat stringent entrance criteria places its dataset somewhere between the PDBbind general set, and the refined set, while Binding MOAD's CSAR-NRC HiQ[64-65] set is more equivalent to the PDBbind core set. At this point in time, PDBbind is the only database that still directly competes with Binding MOAD as a resource of paired binding and structural information.

2.3 sc-PDB

The sc-PDB is a database of ligand-able binding sites based on the PDB, which provides all-atom descriptions of proteins, their ligands, their binding sites and the binding mode for each ligand.[66-67] The 2017 release of sc-PDB contains 16,034 entries, corresponding to 4782 different proteins and 6326 different ligands. Requirements for entry include: Resolution better than 3 Å (NMR structures allowed), valid ligand, and at least one protein chain must have annotations in Uniprot. The sc-PDB aims to be a provider of high quality protein-ligand structures suitable for computational drug design methodology, such as docking. Notably, the sc-PDB does not contain affinity data for their complexes.

The structures in the sc-PDB are protonated, if hydrogens were not in the initial structure, while leaving arginine and lysine nitrogens positively charged as well as aspartic and glutamic acids negatively charged. This protonation is aided by ionized templates built from the HET group dictionary and optimized using the BioSolveIT Hydescorer program.[68-69] Binding sites are constructed using a 6.5 Å heavy-atom-to-heavy-atom cutoff from the ligand, covalently attached ligands and cofactors are allowed. Binding-site water molecules are also identified (any water with at least 2 hydrogen bonds in the binding site) in about two-thirds of the binding sites in the sc-PDB. Some analysis of binding site similarity is also accomplished. One of the primary goals of the sc-PDB is to provide a dataset for inverse docking, specifically to study drug-like molecules.

Due to the pre-processed and refined nature of the sc-PDB, its data set is significantly different than the content of Binding MOAD or PDBbind. This coincides with their mission to provide structures suitable for computational drug design methodology, but they inadvertently remove structural information that would be valuable to the drug design process, too. An otherwise high-quality structure may not have any chains annotated in Uniprot, but still be a newly discovered close homolog to an important target. Such a structure would be granted entrance to Binding MOAD and searching for it would reveal structures and ligands of close homologs also contained in our dataset.

2.4 BioLiP

BioLiP is a ligand-complex centric database maintained by the Zhang lab at the University of Michigan.[70] The focus of this database is to serve a resource for virtual screening and template-based ligand binding site prediction methods. They aim to contain only biologically relevant molecules as ligands in their database, and to expand their collection to also contain DNA/RNA-ligand complexes. BioLiP does address that protein numbering problems with files in the PDB (see 2.7.2 for details) are a community issue and makes some corrected PDB files downloadable. Entrance criterion for BioLiP completely revolve around ligand validity, structure resolution is not considered. BioLiP boasts 407,148 data entries (Accessed 1/2018), with 225,979 regular (non-metallic) ligands and 23,492 binding data. Of these binding data, they attribute 10,971 to be sourced from Binding MOAD, 16,980 from PDBbind, 7331 from BindingDB and 64 from manual curation.

There are two major downsides of BioLiP. Firstly, its 407,148 data entries are not cannon with most of the databases in this field, as each chain of a PDB file is considered a separate entry in BioLiP and therefore the database does not adequately address redundancy when describing its full dataset. Secondly, due to the lack of resolution cutoffs for structure entry, over 150,000 of the database's "entries" are of resolution we would typically consider to be poor (>2.6 Å). While the authors' intended mission of creating a perfectly parseable database for computational purposes is noble, the database is of limited use for our work without major filtering efforts.

2.5 Other Protein-Ligand Databases

PDBbind is the only protein-ligand centric database that operates along the same set of goals as Binding MOAD. Below are some other notable databases that serve different important

functions, as well as a few that are no longer well-maintained. The PDB serves as a centralized resource for structural information for both protein and nucleic acid-based targets, but contains such a breadth of information with such a wide range of quality that external subsets of the PDB's collection are necessary. This being said, the PDB is always making improvements to provide a better website experience, better filtering tools, as well as incorporating more cultivated data from outside resources.

2.5.1 **Binding DB**

The Binding Database (Binding DB) is centered around a high-volume collection of affinity data for small molecule ligands and biopolymers binding to protein targets. [71-72] The majority of this affinity data is in K_i format, but other forms of binding constants are also found. The current version of Binding DB contains 1,427,022 binding data for 639,152 small molecules across 7,026 protein targets. Clustered down, their database reduces to 2291 nonredundant targets at 100% sequence identity, and 5816 targets at 85% sequence identity. Binding DB is also responsible for creation of the datasets for the Drug Design Data Resource (D3R) competition.[73] Binding DB's strength lies in the sheer volume of data encompassing experimental conditions for determination of binding information, even including raw data in some cases. The bane of this resource in the context of this work is that the binding information is rarely coupled to structural information. Trying to condense this information to a reduced number of structures would inevitably result in binding constants being paired to structural data that it does not truly represent due to difference in conditions (i.e. protein crystallized at pH 5 and assay conditions at pH 9). Due to the voluminous direction that Binding DB has taken, it now contends as a resource of the same nature as ChEMBL.[74] Additionally, previous work in our laboratory has revealed that close to 85% of the data in Binding DB is redundant with ChEMBL.[75]

2.5.2 **ChEMBL**

ChEMBL is an extremely large database dedicated to containing binding, functional and ADMET information for drug-like bioactive compounds.[74] Hosted by the European Bioinformatics Institute (EBI) and part of their vast array of publicly available resources, it's a powerhouse for collecting data in the bioinformatics field and BindingDB gets a large amount of its data from ChEMBL.[75] Release version 23 of ChEMBL contains 14,675,320 bioactivity values across 1,735,442 distinct compounds against 11,538 targets. This data was obtained from 67,722

different publications. Importantly, ChEMBL does not necessitate the presence of structural information for its different targets, resulting in many data that do not have a corresponding crystal structure. While ChEMBL is an undoubtable powerhouse for its field of information, as well as rate of content update, culling through the amount of information yielded from this resource is a daunting task and the sections of work in this dissertation all require structural information.

2.5.3 **Relibase, Relibase+**

Relibase, part of the Cambridge Structural Database (CSD), is a dataset collection focused on protein-ligand complexes without affinity information.[76-77] Relibase+ is the premium access version of the database, while Relibase is the web-interface version that is freely available to the academic public. For its initial release in 2002, Relibase+ contained 15,454 PDB entries represented by 50,514 individual ligand sites, with 4530 unique ligands.[76] While this resource is valuable, its definition of valid ligands is far too inclusive, allowing ions, inorganic salts (such as sulfate), and common crystallographic additives such as polyethylene glycol to be considered valid ligands (whereas they are not valid in Binding MOAD). Studying interactions with smaller molecules like ions may be beneficial for studying a specific molecular feature in the context of a specific protein. Binding MOAD considers some of these small moieties to be part of the protein, and ionic species that are not part of a larger molecule are not considered valid ligands at this time, as the database focuses on substrate-like molecules, organic cofactors, and inhibitors.

2.5.4 **LPDB**

The Ligand-Protein Database (LPDB) was an early database with only 195 protein-ligand complexes with binding data representing 51 different receptors across 21 protein classes.[78] LPDB was focused around providing researchers with computer generated docking decoys to aid in developing more accurate scoring functions. This database has been updated very little since its initial release.

2.5.5 **AffinDB**

AffinDB represents a dataset of structural data coupled to affinity data with a number of other experimental details provided, such as SMILES and molecular weight of bound ligands, and pH for the crystallization and biological assay conditions. Published originally in 2006 by Gerhard Klebe and coworkers, AffinDB has not been regularly maintained since that time, containing only

748 affinity data for 474 PDB structures.[79] The small data size of this database leaves it completely redundant or outdated by the content of Binding MOAD and PDBbind.

2.5.6 **Databases: Summary**

PDBbind is the only true competitor to Binding MOAD, providing a similar collection of protein data. The entrance criteria are nearly identical, but the provided subsets of data are where the databases start to differ. The Binding MOAD dataset falls somewhere between PDBbind's general set and refined set, and the HiQ dataset available from Binding MOAD is not as stringent as PDBbind's core set. Neither of these approaches are technically wrong, it all depends on the applications that users are interested in. The sc-PDB is the most similar of the remaining databases, but the pre-processed nature of its dataset puts it into a docking/ *in silico* pre-prep niche that sets itself apart.

ChEMBL and Binding DB provide a tremendous amount of binding data over a significant volume of protein targets, but do not have enough structural data coupled together to classify as the same type of database as MOAD or PDBbind. Relibase is another voluminous database but is gated behind web-access portals which make it somewhat difficult to work with. The gating also makes it difficult to extract details about the contents of its dataset. These difficulties are compounded when Relibase+ is considered, as it is also gated behind a pay-wall. LPDB and AffinDB are two smaller databases that were very similar to Binding MOAD in their initial states, but neither have been regularly updated and are thus no longer as relevant.

2.6 Redundancy

Data redundancy can be a huge issue when discussing large collections of crystallographic data. Many protein complexes have multiple complexes with different bound ligands. Targets with medicinally based biochemical interest such as HIV protease, dihydrofolate reductase, thrombin, and trypsin, as well as targets used to study different biophysical methodologies such as lysozyme and thermolysin, tend to have a very large number of structures in the PDB. Binding databases to-date have rarely addressed redundancy. In the cases that they do, it's usually handled by providing clustered datasets at various sequence identity cutoffs. Binding MOAD's most strict cutoff is 90% sequence identity, which reduces the 25,769 structures to 7,599 unique protein families. In practice, this method of accounting for redundancy usually functions by analyzing each cluster of proteins and selecting one or a few structures that best represent the population. Unfortunately,

this means that some nuances of the collective data are lost, since all of the structures aren't used. We therefore present the idea of extended, or unified binding sites to provide a collective representation of the binding sites in a family of proteins, rather than single depictions thereof.

2.7 Unified Binding Sites

Unified binding sites are a construction of binding sites across an ensemble of structures of the same protein. These extended sites provide a much more robust binding site definition which may illuminate portions of a binding cavity that are only sometimes contacted. This extra data could serve as a powerful tool for *in silico* drug design methodology. Our unified binding sites are constructed utilizing distance cutoff measurements from ligands using coordinates from the corresponding PDB files. Similar conceptual methodology has been presented in the past, primarily by ProBis.[80] Their method of calculating these sites is less straightforward, and extracting the end-information from the various PDB files is difficult because their method does not address the numbering problems described below in 2.7.2.

2.7.1 Construction

Binding sites are typically a representation of protein residue heavy atoms within a given distance cutoff of the heavy atoms of the bound ligand. Commonly, any residue with a heavy atom within 4.5 Å of the bound ligand is considered part of the binding site. Small molecule ligands come in many shapes and sizes, especially in larger binding sites such as protein kinases and proteases, where the natural substrate is very large. Thus, an assembly of many representations of the same binding site may yield a far more robust definition of the binding site.

2.7.2 Protein Numbering

The most difficult aspect of assembling these unified binding sites is making sure that the addressed binding site residues are compatibly formatted. Two structures of the same protein will often be numbered in the same fashion, but it can be exceedingly difficult to identify and fix examples where this is not the case, when using automated scripts for data processing. There are examples of well resolved, high-quality crystal structures that unfortunately suffer from multiple of these numbering ailments. Therefore, we addressed this issue by renumbering protein structures prior to the assembly of unified binding sites to make sure that the merging process was seamless.

A resource of numbering mapping was necessary, as the slew of possible numbering issues made conception of logic-based code to automatically renumber the PDB files exceedingly complicated.

2.7.3 UniProtKB and PDBSWS

The UniProt Knowledgebase (UniProtKB) is a central access point for extensively curated protein information, including details about protein function, classification (including EC information), and even manually annotated cross references to other protein structures.[81] Andrew Martin published PDBSWS in 2005, a resource for cross-referencing protein sequence numbering between PDB structures by cross-mapping with data from UniProtKB.[82] There are a number of fundamental arguments for how protein structures should be numbered, using the carefully annotated data in UniProt resolves this issue. Importantly, the mapping data in PDB SWS is available as a CSV, which is an easily parseable file format.

2.8 Methods

2.8.1 Top-Down Approach

Other protein-ligand databases such as ChEMBL and BindingDB cultivate their data in a "bottom-up" direction, starting with the literature and available binding information for important ligands, and gathering structural data along the way if it is available. Since we are only interested in interactions where corresponding structural data exists, we operate along a "top-down" approach which starts with the PDB. We first import the entire PDB, removed inappropriate structures and use 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, obtaining the appropriate reference for the literature portion of the search is straightforward.

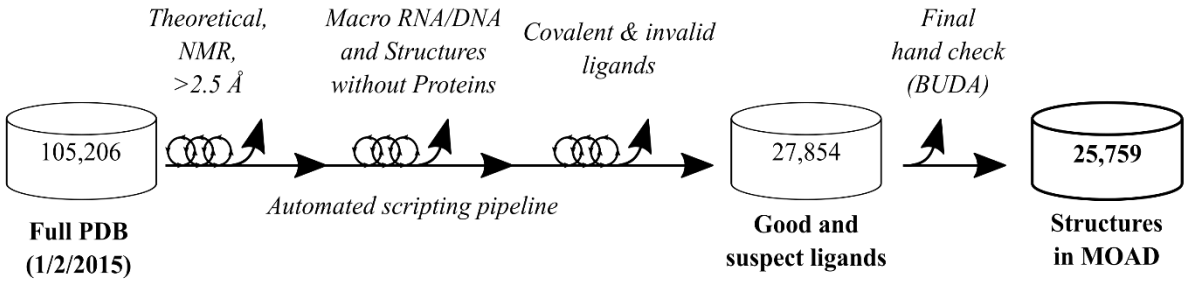
2.8.2 Condensing the PDB

Our data pipeline begins with Perl scripts that assess whether each protein structure is an appropriate entry for Binding MOAD, see Figure 2-1: Binding MOAD Update Process. Our scripts take advantage of the BioPerl toolkit to make parsing PDB files easy.[83] As the original data pipeline of Binding MOAD was developed using mmCIF files, we can accommodate whichever file format is deemed to be up-to-date in the bioscience community.

Starting with the entire PDB (105,206 structures on 1/2/2015), 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. Small molecule ligands were identified in the HETATM and FORMUL sections.

Figure 2-1: Binding MOAD Update Process



Covalently linked ligands were identified by calculating the minimum distance between the protein and each ligand. Minimum distances greater than 2.4 Å were considered noncovalent. Values between 2.1 – 2.4 Å were flagged and examined by hand, in context with the literature to determine covalency appropriately. Distances less than 2.1 Å were considered covalent unless the close contact was to a metal ion (we considered many common catalytic metals to be part of the protein during this analysis). All close contacts to metals were examined visually. This was crucial in the case of zinc-containing enzymes where zinc-ligand distance < 2.1 Å is not necessarily a covalent bond.[84] HET groups within 2 Å of another HET were considered 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 in section 2.7.3), the crystal structure is rejected.

2.8.3 Hand Curation

Literature citations for all final structures to be included in Binding MOAD were read to confirm the validity of the ligands, as well as extract binding data. Our order of preference for affinity data is: $K_d > K_i > IC_{50}$. Great care is taken to ensure that ligands entered into Binding

MOAD are biochemically significant and are of relevant function in the crystal structure being considered. As this initial data screening process is automated, a list of criteria is used to flag ligands into various classifications; Table 2-1 establishes some of the classifications that these ligands may belong to.

Table 2-1: Definitions for Unusual HET Groups

Classification	Type of HET (Examples)
110 Suspect Ligands	**Sugars** (glucose, galactose, fructose, xylose, sucrose, β-D-xylopyranose) **Small organic molecules** (benzene, toluene, phenol, t-butyl alcohol) **Membrane Components** (phosphatidylethanolamine, palmitic acid, decanoic acid) **Small metabolites that may be buffer components** (Citric acid, succinate, tartaric acid)
77 Partial Ligands	**Chemical Groups** (amino group, ethyl group, butyl group, methoxy group, methyl amine) **Inorganic centers of transition state or product mimics** (aluminum fluorides, beryllium fluorides, boronic acids) **Modified amino acids** (oxygens of oxidized CYS, phosphate on TYR)
552 Rejected Ligands	**Unknown or dummy groups** (UNK, DUM, unknown nucleic acids or fragments thereof) **Salts and buffers** (Na^+, K^+, Cl^-, PO_4^{3-}, CHAPS, TRIS, Me_4N^+) **Solvents** (DMSO, hexane, acetone, H_2O_2) **Crystal additives, cryoprotectants, and detergents** (Polyethylene glycol, oxtoxynol-10, dodecyl sulfate, methyl paraben, 2,3-propanediol, pentaethylene glycol, cibacron blue) **Metal complexes 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+}) **Catalyic centers that are part of the protein** (4Fe-4S cluster, Ni-Fe active center) **Heme groups** (heme D, bacteriochlorophyll, cobalamin, protoporphyrin IX)

For brevity, not all compounds are listed in this table.

Many factors aside from the identity of the ligand are also considered. Short protein-ligand distances and suspected ligands are flagged for manual inspection during the hand-check stage. Suspect ligands are typically crystallographic additives but may be valid ligands in some cases. Partial ligands are unlikely to be stand-alone entities and typically represent a portion of a multi-part ligands. Any HET with 3 heavy atoms or fewer is automatically part of this list. The covalency check identifies if these HET groups are modifications to the protein or a ligand.

Modifications to amino acids are on the partial ligand list because they can be part of the protein or part of a peptide ligand, and their listing in PDB files varies in both cases. Complexes containing heme groups must contain another valid ligand, as heme groups are considered to be part of the protein and not a ligand. Due to the close proximity of various metals and even small molecule ligands to heme groups, determining valid ligands in these cases can be convoluted. We refer to the corresponding literature for clarification whenever possible. Due to this nature, many cytochromes are excluded from Binding MOAD, which we acknowledge are very useful targets to have biochemical information for. We plan to add support for more robust description of these enzymes to Binding MOAD in the future.

2.8.4 **Addressing Redundancy by Sequence**

Grouping proteins by similar sequence allows users to find multiple related structures, which makes various types of comparison and dataset construction much easier. Enzyme classification (EC) numbers are used to group enzymes that perform similar catalytic reactions. In the past, Binding MOAD clustering was based on EC groupings, but this method has been abandoned for a number of reasons. The EC number listed in PDB files is not always correct, or present at all. In the latter case, filling in the missing data gaps is convoluted. But, most importantly, there still exists massive variation within Enzyme Classifications, so grouping into homologous protein families by sequence has proven to be more beneficial, straightforward, and reproducible. Structure sequences are compared using BLAST[85] and family leaders are chosen as detailed in this schema:

1. Use BLAST to compare each protein chain of each entry to all other chains, different sequence cutoffs are used for a few different perspectives:
    a. 90% sequence identity
    b. 70% sequence identity
    c. 50% sequence identity

2. Choosing the family leader: The details of this priority list are as follows:
    a. Tightest binder (where binding data is available)
        i. 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.
    b. Best resolution (complexes with ligands preferred over cofactor-only complexes)
    c. Wild-type over structures with site mutations
    d. Most recent deposition date
    e. Factors such as R or R_{free} values
    f. If all the above criteria are identical, the entries are likely from the same paper, which will be used to help in the tie-breaker

2.8.5 **Addressing Redundancy using Unified Binding Sites**

The most difficult aspect of assembling unified binding sites comes in the form of protein numbering. Protein numbering is rarely always done the same way in a group of more than a few structures. This is emphasized in a number of somewhat common cases:

- Numbers are simply skipped over, attempting to relate a given protein structure to a relative in an enzymatic cascade where perhaps a fragment was removed mid-sequence in a previous cascade transformation
- Awkward nomenclature for starting new protein chains:
    - Adding some base number to the residue number to encode the chain: Chain A: 1001, 1002, 1003, 1004… ; Chain B: 2001, 2002, 2003, 2004…
    - Picking up where you "left off": Chain A: 303, 304, 305, TER 305 ; Chain B 306, 307, 308
- Residual purification tags on the N-terminal ends of the protein that are not accurately accounted for in the SEQRES section of the PDB file
- Unresolved residues are not always accounted for in the protein numbering, sometimes the number gap is left and sometimes it is skipped
- Insertions are not always numbered as they should be (3a, 3b, 3c, 3d vs. 3, 4, 5, 6)

    - This also applies to point 1. Where it could be attempting to represent the biological relative proteins, where perhaps a sequence fragment is added during an enzymatic cascade
- Largely unresolved loops of protein with repeat amino acids in the sequence, where only one of them is resolved (i.e. a sequence of Arg Lys Lys Glu, with one Lys as the only residue resolved).

Two structures of the same protein will often be numbered in the same fashion, but as more data is added from more different sources, it becomes increasingly likely that this will not be the case. It is exceedingly difficult to identify and fix examples where numbering issues arise when using automated scripts for data processing. There are many examples of well resolved, high-quality crystal structures that unfortunately suffer from multiple of these numbering ailments. Therefore, we addressed this issue by renumbering protein structures prior to the assembly of unified binding sites to make sure that the merging process was seamless. A resource of numbering mapping was necessary, as the slew of possible numbering issues made conception of logic-based code to automatically renumber the PDB files too complicated.

To start, a similarity matrix of all protein chains in Binding MOAD was constructed, using sequence alignment tools; this was accomplished with both NEEDLE[86] (part of the EMBL-EBI toolkit) and with BLAST at different times.[85, 87] PDB SWS was used as our resource for renumbering templates.[82] PDB structures were then renumbered using the following framework:

1. If the PDBid/chain combo is found in PDB SWS, renumber it accordingly
2. If the PDBid is found in PDB SWS, but not for the current chain, use any sequence identical chain within the same PDBid that is found in PDB SWS
3. If the PDBid is not found in PDB SWS, use another structure that has a 100% sequence identical chain as the renumbering template.
4. If no structures in a homologous family are found in PDB SWS, check to see if their numbering already matches up
    a. These cases are usually small homologous families, where manual inspection of this type is reasonable

In cases where multiple renumbering frameworks were provided by PDB SWS for a single homologous family, the mapping for the family leader was chosen and the whole family was renumbered in the same manner.

2.8.6 **Annual Updates**

We conduct annual updates to incorporate more structures into Binding MOAD as they become available from the PDB. Our 2017 update began in January. The update procedure follows:

1. Use the PDB's list of obsolete entries to identify any existing structures in Binding MOAD that should be removed or replaced.
2. Download new set of PDB files. The previous version will be compared to identify new structures that have been added to the PDB since the last update.
3. Identify good protein-ligand complexes in the new structures using our script pipeline.
4. Script outputs are then double checked by hand to validate the outputs. No new structures are entered into MOAD without manual inspection.
    a. New het groups are examined by hand during this process
5. Sequences are re-binned into new homologous family groupings, and new leaders are chosen as described in 2.6.5.

2.9 Results and Discussion

The creation and maintenance of Binding MOAD is an assembly of many years of work by several different people. I am directly responsible for developing the data pipeline for renumbering PDB files using PDB SWS and subsequently determining unified binding sites.

I also dedicated a significant amount of time towards the annual updating of Binding MOAD's dataset. Between the 2010-2014 releases of Binding MOAD, ~8800 new structures were added, and I was responsible for approximately 3500 of those structures. These updates take a considerable amount of time to complete, even with multiple people. The advent of PDF availability and E-publication was a significant boon to this process but has also aided the PDB's rate of growth. Using keyword searches is advantageous for text-rich papers, some examples of important key words to search include: 'ki', 'kd', 'ic50', 'affinity', 'bind(ing)', 'association', 'dissociation', 'constant', 'inhibitor', 'covalent', 'coordinate', and 'inhibition'. Some articles contain crucial information in their figures or captions, but not all figure captions and figures are easily parseable with 'find' functions.

The most recent update of Binding MOAD was derived from the version of the PDB extracted on January 2, 2015 (105,206 entries), a total of 25,759 valid protein-ligand complexes were obtained. Binding MOAD contains 12,432 unique, valid ligands within the 25,759 complexes. Figure 2-2 provides the distribution of these valid ligands by molecular weight. The ligands range from 4-278 heavy atoms, with an average molecular weight of 433 g/mol; an example of the average ligand is adenosine-5'-diphosphate (ADP), which has a molecular weight of 427 g/mol. Figure 2-2 shows that the number of large ligands (>500 g/mol) drops off quickly. The largest ligands are sugar, peptide, and nucleic acid chains.

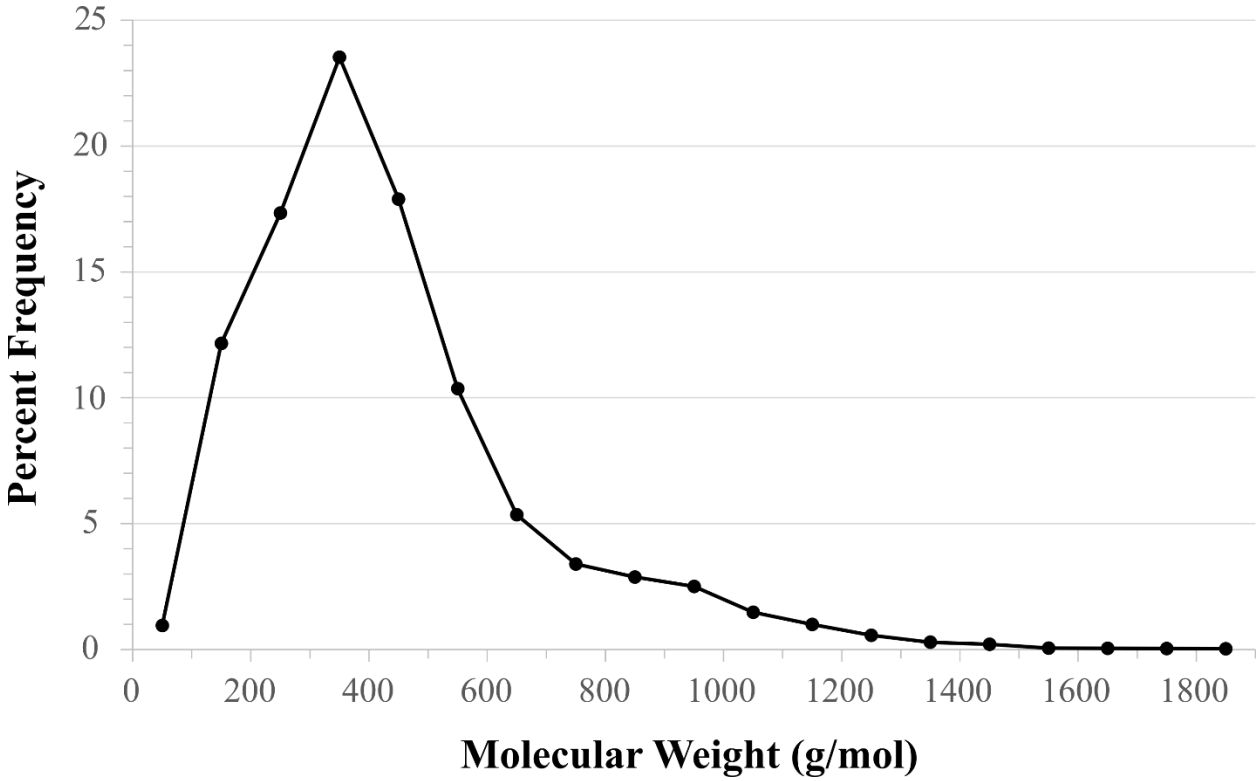


Figure 2-2: Distribution of the current 12,432 unique ligands by molecular weight.
The average ligand in Binding MOAD is 433 g/mol. The largest are polysaccharides, peptides, and polynucleic acids.

Binding MOAD also contains 9138 binding data across the 25,759 complexes. These binding data are composed of 2937 K_d or K_a, 3104 K_i, and 3097 IC_{50} values. These binding affinities range over 16 orders of magnitude; Table 2-2 presents median, tightest, and weakest binding values for each type of binding data, and the distribution of the three types of binding is presented in Figure 2-3. For Figure 2-3, the binding data are represented as free energy of binding (-kcal/mol), using -RT *ln* (data), where the values were converted to molar units and K_a data were

converted to K_d. This application is simply for ease of viewing and comparison for the reader, as the assumption of 298 K temperature is not strictly appropriate for many values, and this approximation is not always relevant for K_i and IC_{50} experimental conditions for other reasons.

Table 2-2: Average and Median of binding data within Binding MOAD

Classification	Median	Tightest	Weakest
K_d, K_a (as $1/K_a$)	**2.48 uM**	**10 pM**	**1.4 M**
K_i	**319 nM**	**11 pM**	**0.837 M**
IC_{50}	**148 nM**	**12 fM**	**0.355 M**

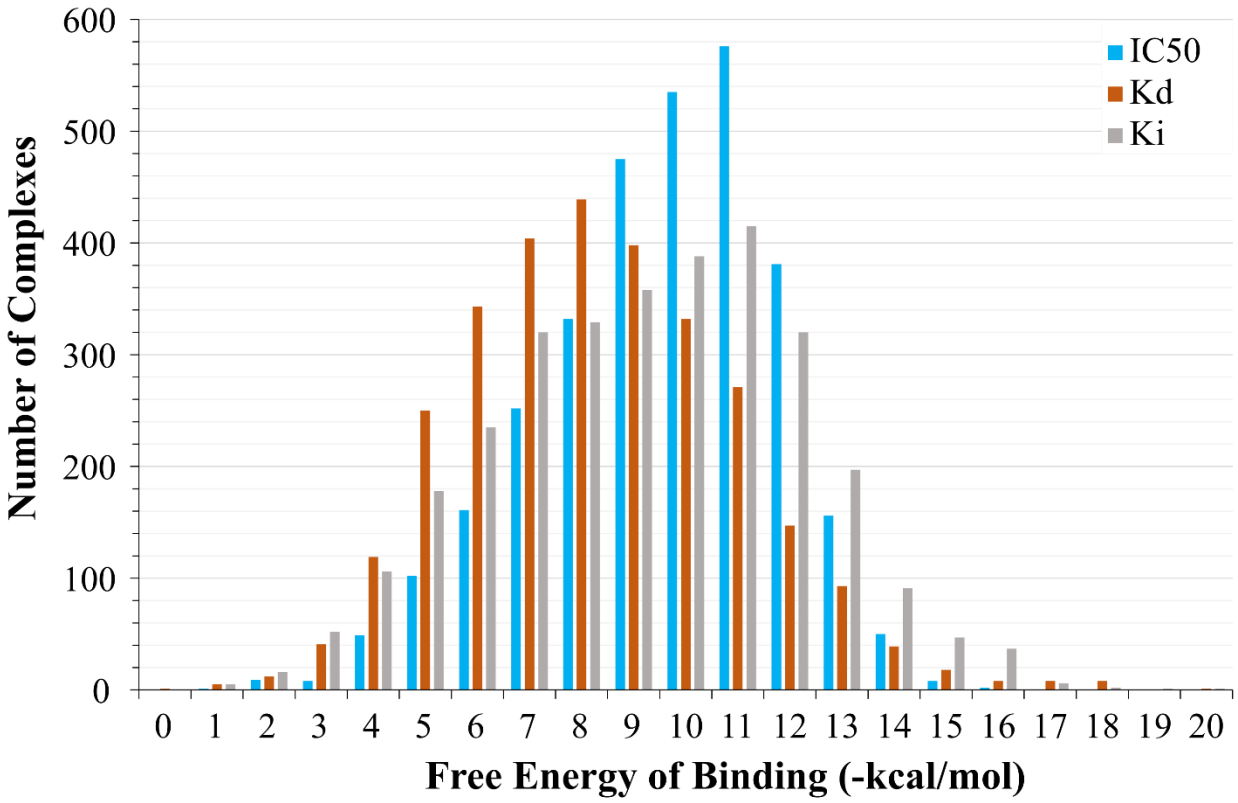


Figure 2-3: The distribution of binding-affinity data within Binding MOAD.
Data is available as K_d (orange), K_i (grey), or IC_{50} (blue). For this histogram, binding data were converted to free energies by -RT ln (data), where the data were converted to molar units and K_a values were converted to K_d.

2.9.1 Clustering Binding MOAD into Homologous Protein Families

The protein sequences of the entries in Binding MOAD were grouped into homologous protein families. When clustered at 100% sequence identity, 13,539 unique protein sequences were identified. When the criterion for sequence identity is relaxed, fewer protein families are found and the size of those families increases, which is expected. Clustering at 90% sequence identity results in 7599 protein families, which is our preferred way to portray the dataset, 70% sequence identity yields 6348 families, and 50% yields 4913 families. Applications of Binding MOAD typically do not utilize the 100% sequence identity families, but this extremely stringent grouping is utilized heavily in the creation of the protein flexibility dataset used in Chapter 3 and Chapter 4.

2.9.2 Database Growth and Updates

As mentioned previously, we are committed to the growth of Binding MOAD as a quality data resource in the community. Since being introduced in 2004, Binding MOAD has regularly expanded its collection with new data. Early updates brought in ~1500 new structures each year, but the rapid growth of the PDB has afforded us with many more structures in recent years. The growth of Binding MOAD is presented in Table 2-3.[3]

Table 2-3: Growth Data for Binding MOAD (2004-2014)

Release (version, PDB download date)	**Protein-ligand complexes**	**Protein Families**	**Unique ligands**	**Binding affinity coverage**
Initial release in 2004[1]	5331	1780	2630	1375 (25.8%)
Prior to website in 2005	8250	2732	3932	2374 (28.8%)
1st (v2006, 12/31/2006)[2]	9836	3151	4665	2950 (30.0%)
2nd (v2007, 12/31/2007)	11,366	3583	5348	3452 (30.4%)
3rd (v2008, 12/31/2008)	13,138	4078	6210	4146 (31.6%)
4th (v2009, 12/31/2009)	14,720	4624	7064	4782 (32.5%)
5th (v2010, 12/31/2010)	16,948	5198	8140	5630 (33.2%)
6th (v2011, 12/31/2011)	18,764	5772	9048	6311 (33.6%)
7th (v2012, 12/31/2012)	21,109	6443	10,156	7284 (34.5%)
8th (v2013, 12/31/2013)	23,269	6960	11,173	8156 (35.0%)
9th (v2014, 01/02/2015)	25,759	7599	12,432	9138 (35.5%)

2.10 Conclusions

We have detailed the further development and are continuing to expand Binding MOAD. In the future, we aim to continue our annual updates to keep pace with the growth of the PDB. Binding MOAD has over twenty-five thousand, hand-curated, protein-ligand X-ray crystal structures that contain ligands of biological relevance. Binding data is available for over one-third of the entries, and this coverage has only increased with every update of the database. The value of Binding MOAD is not necessarily present in the quantity of its data, but more-so in the quality. Maintaining this data quality is only achievable due to the considerable amount of effort placed in the update process and hand-curation. We are planning to add similarity-based metrics to search the dataset, both in terms of ligand similarity as well as protein similarity. Furthermore, we plan to incorporate more benefits of Natural Language Processing (NLP) into our data curation pipeline, to help streamline the process and reduce the heavily curator-intensive nature of the database updates.

Our datasets are available online at http://www.BindingMOAD.org. This web-accessible resource is available to the research community, and our web interface also allows for users to contact us if they find any aspects of our curated data to be incorrect. Each structure's webpage includes: Details about ligands (both valid and invalid), available binding data, PDBid for structural coordinates, EC class, homologous protein families with links to related structures at multiple sequence cutoffs (90%, 70%, 50%), as well as a protein viewer for visualization of the ligand bound in the extended binding site (using NGL viewer).[88] When searching by PDBid, users may be presented with information as to why a particular structure was excluded from Binding MOAD's dataset (resolution > 2.5 Å, no appropriate ligand, etc.).

Downloadable copies of our dataset are available from the download page of our web server. We have chosen to make the structures available as biological unit files as opposed to PDB files. Biological units contain the proper multimer for biological activity, as opposed to PDB files which do not necessarily do so. Proteins that occur in formations with easily divisible symmetry operators, such as dimers, are often misrepresented in PDB files, where the only true dimeric representation of that protein occurs when a dimeric pair crystallizes in the same unit cell. Utilizing biological unit files provides users with structures that are most related to the biological activities listed in Binding MOAD, and thus the best representation of our mission to provide high quality structural information complimented by appropriate binding data.

Chapter 3. Protein Flexibility and Ligand Binding

3.1 Abstract

Understanding how ligand binding influences protein flexibility is important, especially in rational drug design. Protein flexibility upon ligand binding is analyzed herein using 305 proteins with 2369 crystal structures with ligands (holo) and 1679 without (apo). Each protein has at least two apo and two holo structures for analysis. The inherent variation in structures with and without ligands is first established as a baseline. This baseline is then compared to the change in conformation in going from the apo to holo states to probe induced flexibility. The inherent backbone flexibility across the apo structures is roughly the same as the variation across holo structures. The induced backbone flexibility across apo-holo pairs is larger than that of the apo or holo states, but the increase in RMSD is less than 0.5Å. Analysis of χ_1 angles revealed a distinctly different pattern with significant influences seen for ligand binding on side-chain conformations in the binding site. Within the apo and holo states themselves, the variation of the χ_1 angles is the same. However, the data combining both apo and holo states show significant displacements. Upon ligand binding, χ_1 angles are pushed to new orientations outside the range seen in the apo states. Influences on binding site variation could not be easily attributed to features such as ligand size or X-ray structure resolution. By combining these findings, we find that binding site flexibility is compatible with the common practice in flexible docking, where backbones are kept rigid and side chains are allowed some degree of flexibility.

3.2 Introduction

Proteins are naturally flexible biopolymers composed of a string of amino acids folded into a largely non-covalent structure.[89] The degree of flexibility is often tightly coupled to the protein's function, especially for enzymes. Understanding the flexibility in proteins is important in 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.[8-9] Rigidity can aid in specificity and tightness of ligand binding, while flexibility allows for entry of ligands into the binding site and can also be involved in communication between allosteric and orthosteric binding sites. Clusters of residues near binding sites are often observed in strained conformations.[6-7] 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 a ligand from an active site.

Being able to fully account for induced changes is especially important in protein-ligand docking. Docking proves to be very difficult in practice when conformational changes occur upon binding.[90-91] 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.[92-95] The larger the required adjustment, the harder it is to accurately predict protein-ligand binding.[96] Protein flexibility needs to be incorporated to accurately represent protein-ligand binding.

As we outline below, there have been many studies examining the extent and properties of ligand binding by comparing apo and holo protein crystal structures. A number of studies have also examined the local characteristics of their binding sites, such as side-chain flexibility or solvent accessible surface area (SASA), while some studies have examined only global protein changes upon ligand binding. Analyses of most studies fell into two categories: root mean square deviation (RMSD) calculations of backbone atoms or rotameric analysis of amino acid side chains. These different approaches have led to conflicting conclusions which our study helps to reconcile. Below, we summarize the most significant findings to date.

3.2.1 Backbone Analysis

Structural variation appears small when assessed through backbone motion. Gutteridge and Thornton found that enzymes in their small dataset of 11 proteins (11 apo, 14 holo) 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).[97]

Gutteridge and Thornton followed their work noted above by looking for conformational changes upon ligand binding in a larger set of structures. In their study of 60 enzymes, ~75% of holo-apo pairs had C_α RMSD of ≤ 1 Å. This RMSD was contrasted with the C_α RMSD observed among apo-apo protein pairs as a baseline, where ~83% of 31 apo-apo pairs had a C_α RMSD of ≤ 1 Å.[98]

Gunasekaran and Nussinov classified 98 proteins into three categories based on maximum C_α displacement between holo and apo structures: rigid proteins (≤ 0.5 Å), moderate (0.5 Å < and ≤ 2.0 Å), and flexible (> 2 Å).[99] 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 and ψ changes between apo and holo were minimal. All classes had a few binding site residues with φ and 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. Furthermore, they found no notable difference in SASA of the binding site residues of their three classifications of binding sites (rigid, moderately-flexible, and very-flexible).[99]

Brylinski and Skolnick found 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.[100] For 521 single-domain apo-holo structural pairs, 80% had an RMSD ≤ 1 Å, and among a set of single-domain holo-holo pairs, ~ 92% had an RMSD ≤ 1 Å.

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, binding site RMSD changes ≤ 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.[101] However, they found that these small movements were capable of inducing significant changes in volume and shape such that volume similarity indices (η) ranged from 0.44 to 0.90. The disparity in geometric similarity indices point to the need for other modes of analysis to accompany RMSD. These results hint that small changes in backbone displacement can result in greatly increased availability of side-chain conformational space.

3.2.2 **Side Chain Analysis**

Analysis of side chains reveals additional qualities of protein flexibility and highlights the detriment of excluding side-chain motion 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.[102-104] They compared their flexible 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 ≤ 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° and 82% of the side chains less than 15°. This range of movement used in SLIDE is comparable to 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°, and 75% changed by < 15°. Thus, small changes are typical, but more importantly, they are critical for accurate results in half of their studied protein structures.

Heringa and Argos have also described how ligand binding was sufficient to induce strain and push some binding site side chains into rotamers outside of the typical minima.[6-7] This encourages the idea of rotameric changes being heavily influenced by ligand binding events.

Zhao, Goodsell, and Olson examined flexibility differences between amino acids.[105] They examined the variation of χ_1 angles among different apo structures of the same protein to establish limits of natural variation in the 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.[106] 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 in binding-site residue side chains are significant enough that most rigid docking will fail.[96, 104, 107] 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.

Najmanovich and coworkers further explored side-chain flexibility utilizing their SEQ dataset, which contains 188 apo-holo protein pairs.[108] They concluded that at least one residue in the binding site undergoes significant rotameric change upon ligand binding in about 88% of their tested cases. At most, five rotamer changes account for all observed movements in 90% of their test cases, and rotamer changes are essential in 32% of flexible binding sites. The different amino acids were shown to have an 11-fold difference in their probability to undergo changes. There are two major takeaways from this work. First, at least one flexible residue is present in nearly all of the binding sites that they tested. Second, different amino acids have notably different propensities to undergo change in rotameric confirmation.

Current Study

The previous studies reveal that there are often a few key residues with significant flexibility within binding sites of otherwise rigid residues that do not undergo significant rearrangement upon ligand binding. However, some of the studies noted above are limited to very small sets of proteins. Additionally, none of the studies covered all three comparison types: Apo-apo, holo-holo, and apo-holo.[109] This is especially important because analysis of induced flexibility (apo-holo) has little relevance without first knowing the inherent variability in each structure type (apo-apo, holo-holo). While changes from ligand binding have been observed, they have not been appropriately separated from inherent variation in proteins.

This study aims to assess protein flexibility upon ligand binding, employing a large dataset and focusing on contrasting inherent flexibility to changes upon binding. Each protein in the dataset has at least two holo and two apo structures, so we may compare the observed variation observed in proteins with ligands (holo-holo pairs), without ligands (apo-apo pairs), and between the two sets (apo-holo pairs). We use a large and carefully created dataset so that the observed differences can be statistically quantified. This study describes a comprehensive set of 305 protein sequences, represented by 2369 holo and 1679 apo protein crystal structures. We describe statistically significant differences in flexibility upon ligand binding. To confirm these changes are

truly due to flexibility, correlations to other properties such as ligand size and crystal-structure resolution were investigated.

3.3 Methods

3.3.1 Holo Dataset Curation

The non-redundant holo structure dataset was derived from Binding MOAD, a source of high quality protein-ligand complexes that have a maximum of 2.5 Å resolution.[3] Biologically relevant ligands are differentiated from opportunistic binders in the crystal structures (e.g. salts, buffers, phosphate ions) of Binding MOAD, making curation of relevant ligand structures straightforward. Furthermore, use of Binding MOAD excludes covalently attached ligands. Structures with more than one valid ligand were excluded from this study in favor of binary protein-ligand complexes to ensure that only one pocket was being analyzed in each protein. Any structures containing additional molecules in their binding site, such as additives, were also excluded.

Each structure in the holo set was clustered based on the sequence identity using stringent criteria of 100% sequence identity in both directions. A subsequent 95% sequence identity clustering of those families was then performed to suggest any families that should be merged due to simple N or C terminal amino acid additions. Sequence identity between structures was determined using BLAST.[110] Any families differing in protein core sequence were kept separate. An index of all apo and holo structures for each protein family in this dataset is provided in Appendix A.

3.3.2 Apo Dataset Curation

A set of apo structures was first compiled by screening the PDB for structures of 2.5 Å resolution or better and then identifying only the structures without any HET groups (except for water) or only having HET groups that are not biologically relevant (such as crystallographic additives).[58] Acceptable additives were restricted to HET groups of 5 atoms or less and 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 dataset at this point. An index of all apo and holo structures for each protein family in this dataset is provided in Appendix A.

3.3.3 File Setup and Preparation

These steps were taken prior to any calculations. The first biounit model containing the relevant ligand of the corresponding PDB structure was used by default for each structure. All hydrogens were removed from the files. Ligand data was extracted, and then all ligands were removed from the files. Waters were removed.

All protein systems were renumbered utilizing the pdbSWS prior to binding site calculation and assembly.[82] In the cases where this would result in more than one numbering pattern inside of a family, one structure's numbering was applied to the other structures. If this was not possible and there was no method to renumber a structure to the same pattern as the rest of its family apart from manual processing, it was discarded from the dataset out of consideration for reproducibility.

Renumbering structures was necessary because some structures were numbered differently (especially common when going between apo and holo structures). Protein numbering becomes critically important in the case of unified binding sites, where it is necessary to harvest residue data from the site when there are no ligands present to define the site (apo structures).

3.3.4 Ligand Size

The molecular weight for each ligand was extracted from Binding MOAD. A unique feature of this set is the size of the ligands involved. This dataset allows for ligands composed of more than one HETATM group from the crystal structure. This study allowed peptides up to 10 amino acids, nucleotides up to 4 nucleic acids, and other multi-HET ligands. Multi-HET ligands were appropriately treated as one large molecule. For example, the inhibitor Aeruginosin98-B, in the PDB structure 1AQ7 of bovine trypsin, is comprised of the HET groups "34H+DIL+XPR+AG2". Newer HET groups have been made recently that combine some of these multipart ligands, but these were not yet implemented to the PDB at the time of this analysis.

3.3.5 Binding Site Identification and Compilation of the "Union" Binding Sites

Each site was defined to include all protein residues within 4.5 Å of the biologically relevant ligands, which should capture both hydrogen-bonding and van der Waals interactions. Hydrogen atoms were not considered in the distance calculation for either the protein or the ligand.

Most of the crystal structures for a given protein had different ligands bound, so many could have a slightly different set of residues near the ligand. Therefore, the summation of all sets of residues in all complexes for each protein was used to identify the "union" binding pocket for that protein, or its unified binding site.

3.3.6 Maximum RMSD and χ-angle Range Calculations

In order to compare the overall similarity of all the structures of a protein, we calculated a maximum RMSD. RMSDs are pair-wise comparisons, and our analysis compared all structures of the same protein to one another. The RMSD calculations were based on all Cα in the backbone of the protein. Methods established previously in our lab were used to compute standard RMSD.[111] Binding-site RMSD values were also calculated for the unified binding sites.

To examine the flexibility of the side chains, χ_1 was measured for residues, except Gly and Ala, utilizing an in-house Perl script. Pro was measured as a control for method proofing, but it is not shown in any of the presented data. Valines are represented by both of the two available χ_1 angles, as Valine is symmetrical at atom γ for the calculation. Isoleucine also contains two χ_1 angles, but the angle used in this analysis is that of the longer carbon chain.

Binning for χ_1 angle plots was accomplished *via* an in-house Perl script. Data for each amino acid was binned on a per-residue, per-family basis, and then averaged over the total number of that residue in the entire dataset. For example, there are 405 Arg residues in the 305 binding sites, each of those are binned with their corresponding data, and then the bins for each of the 405 residues are averaged together to represent all Arg residues in the dataset.

The variation for a given residue was measured by determining the *range* of χ_1 values observed for each residue in each binding site. This range is the smallest mathematical angle that contains all χ_1 angles observed for each amino acid. (eg. Values of 30˚, 45˚, and 100˚ would yield a χ_1 angle range of 100-30 = 70˚.)

3.3.7 SASA Calculations

SASA was calculated using NACCESS.[112] Default probe size (r = 1.4 Å) was used. All hydrogens, ligands, water, and HET groups were removed prior to calculation. This is default behavior for HET groups, however in the case of peptide ligands it is necessary to remove the ligand from the file. SASA was calculated for all residues of the protein sequences first, and then the binding sites were extracted for analysis.

3.3.8 Statistical Methods

To assess if an observed difference between two groups (being apo, holo, or the entire apo-holo dataset) is sufficient to reject the null hypothesis that the two groups have identical distributions, Wilcoxon signed-rank tests were used (performed in JMP).[113] These tests applied to the distribution of maximum RMSD measurements among the families and average χ_1 angles.

In lieu of that manner of statistical test, error bars describing 95% confidence intervals for Figure 3-6 were jack knifed by 1000 resamples of the data at 90% of the original dataset's size or "leaving 10% out."

All statistical analyses were performed utilizing the statistical packages JMP and R.[113-114]

3.4 Results and Discussion

3.4.1 Dataset Properties

The most recent release of Binding MOAD[3] was clustered to obtain relevant holo structures, matching apo structures were obtained from the RCSB Protein Data Bank (PDB)[58-59] as described in the methods section. Upon filtering for proteins with at least 2 holo structures and 2 apo structures, this dataset reduces to 305 different proteins, represented by 2369 holo structures and 1679 apo structures. An index of all apo and holo structures for each protein family in this dataset is provided in Appendix A. Our dataset is over an order of magnitude larger than the previously utilized datasets for this type of study. Our dataset has relatively low redundancy with previously utilized datasets. For example, Skolnick's dataset of 521 apo-holo protein pairs has only 69 apo and 49 holo structures in common with our dataset.[100]

The proteins with the most holo structures are carbonic anhydrase II followed by trypsin, with 174 and 120 holo structures, respectively. The proteins with the most apo structures are lysozyme followed by ribonuclease-A, which have 280 and 79 apo structures, respectively. This redundancy is accounted for by giving each protein family one overall value (a maximum, mean, or median) to describe all of its data in our analyses.

The ligands in our dataset are diverse and represent many different classes of molecules. The average molecular weight of the ligands is 374 g/mol with 80% of ligands less than 500 g/mol and 95% less than 800 g/mol. The heaviest ligand is a seven-residue synthetic peptide bound to

Endothiapepsin in structure 4LP9 at 999.18 g/mol. The smallest ligand is hydantoin which is bound to a Zinc dihydropyrimidinase in crystal structure 4LCS and weighs 100 g/mol.

The crystal structures of the apo proteins and holo proteins have average resolutions of 1.82 Å and 1.84 Å, respectively. Apo and holo median resolutions were 1.84 Å and 1.85 Å, respectively. There were 159 families better resolved in their holo form, 139 families better resolved in their apo form, and seven families with the same average resolution in both their holo and apo forms. Therefore, there is little bias between the resolution of the holo and apo sets that could influence the measurements used in this study.

3.4.2 Unified Binding Sites

Traditional descriptions of ligand binding sites use residues within some distance cutoff of the ligand contained in a protein crystal structure. Contacts here are defined by a 4.5 Å cutoff between any heavy atom of a bound ligand, and any heavy atom belonging to an amino acid residue in the protein sequence. Our "unified binding sites" are a union of all residues within a 4.5 Å distance cutoff from any bound ligand within any of the holo structures in a protein family (hydrogens were not considered). These unified binding sites represent the totality of the binding site. Union binding sites averaged 21 ± 9 amino acids in size.

3.4.3 Flexibility of Protein Backbones

Backbone RMSD overlays for the entire backbone of all structures of each protein were obtained and the maximum RMSD value for each type of pairing (e.g. apo-apo, apo-holo, holo-holo) within each family was determined (see Methods). Family maxima were chosen instead of medians or averages to readily identify proteins capable of large conformational changes. Table 3-1 presents the averages and medians of the maximum RMSD values for the 305 unique proteins. Distributions of the maximum RMSD are given in Figure 3-1. The maximum RMSD for the apo pairs, holo pairs, and apo-holo pairs are compared for each of the protein families in Figure 3-2.

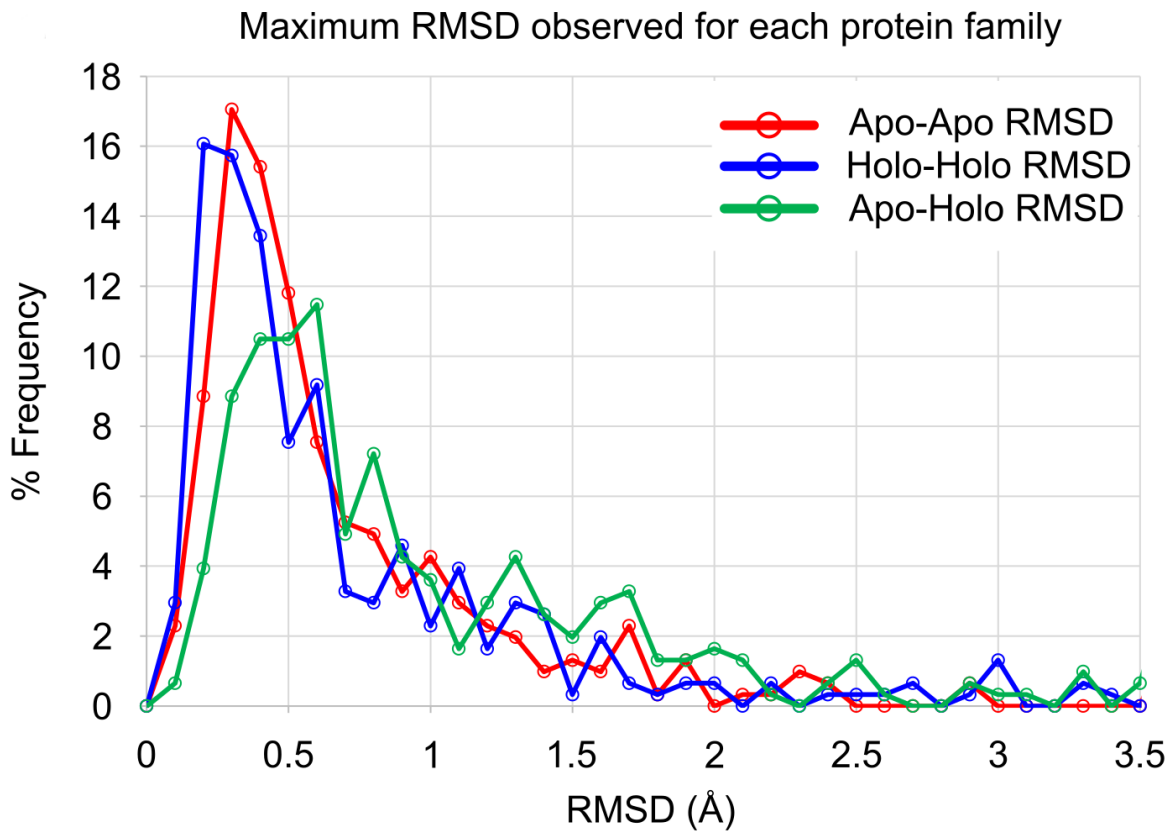


Figure 3-1. Distribution of maximum backbone RMSD for each protein family. The data for the apo-apo pairs is shown in red, holo-holo pairs are shown in blue, and apo-holo pairs are shown in green. There is no statistical significance to the difference in apo-apo vs holo-holo data ($p > 0.05$, difference in medians = 0.025 Å). The difference between the apo-holo data and apo-apo data are significant ($p < 0.0001$, difference in medians 0.241 Å), as is the difference between the apo-holo and holo-holo data ($p < 0.0001$, difference in medians 0.266 Å).

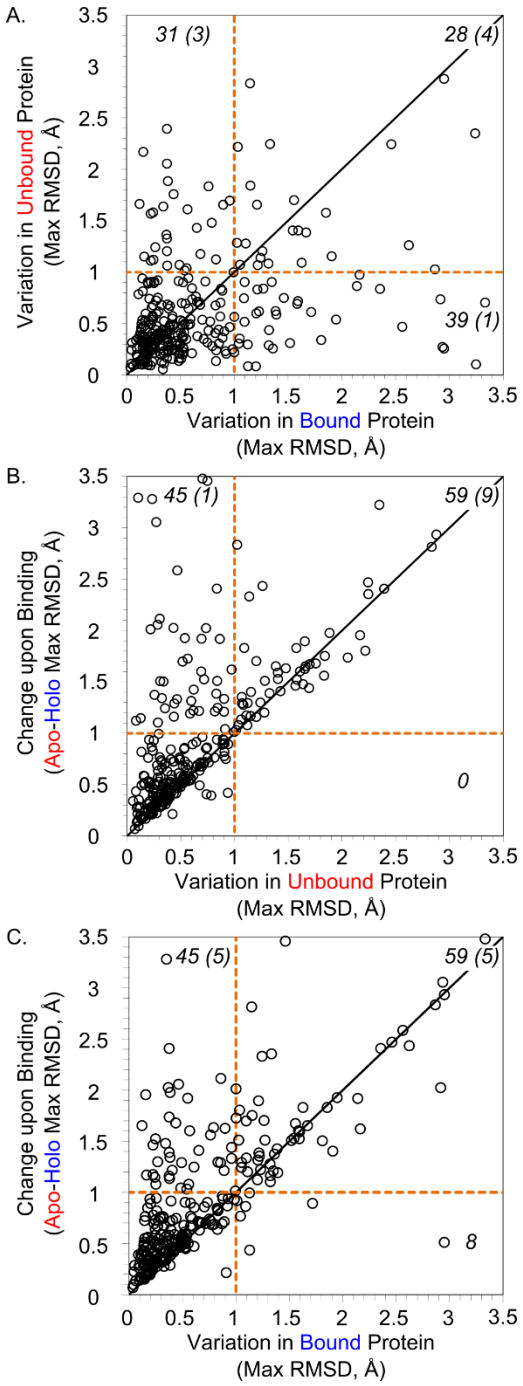


Figure 3-2. Analyses of maximum backbone RMSD for each protein family. Each point represents the maxima observed in one protein family, and the number of points of each section is labeled in black (numbers in parenthesis are points with values > 3.5 Å). A) The maximum across the apo-apo pairs is compared to the maximum of the holo-holo pairs; 207 proteins display RMSD ≤ 1 Å for both groups. B) The maximum across the apo-holo pairs is compared to the maximum of the apo-apo pairs; 201 proteins display RMSD ≤ 1 Å for both groups. C) The maximum across the apo-holo pairs is compared to the maximum of the holo-holo pairs; 201 proteins display RMSD ≤ 1 Å for both groups.

Apo structures and holo structures have similar conformational variation based on the comparison of the maximum apo RMSDs versus maximum holo RMSDs of each protein (Figure 3-1 and Figure 3-2A). In general, proteins tend to have the same conformational flexibility within the apo and holo states. Only 10% of the proteins' apo structures show significantly greater backbone flexibility than their holo structure counterparts, and 12% of the proteins' holo structures show significantly greater backbone flexibility than their apo structure counterparts (31 apo families, 39 holo families). There were 28 families with both Apo and Holo maximum RMSD > 1 Å, indicating that both binding states are relatively flexible. The maximum backbone RMSD for apo and holo structures were both < 1 Å for 207 of the 305 proteins, showing that ~68% had negligible conformational flexibility regardless of ligand binding. Wilcoxon signed-rank tests support this, showing that apo vs. holo data distributions are not significantly different with $p > 0.05$ (see Methods).

As we expect, there is greater variation seen in going between apo-holo pairs (Figure 3-1 and Figure 3-2B-C). Compared to apo-apo and holo-holo pairs, 15% of proteins (45 protein families) have significantly more conformational space available to their backbones between the unbound and the bound state (apo-holo pairs) when compared to either the apo (Figure 3-2B) or holo (Figure 3-2C) states. Importantly, these 45 protein families are not completely redundant between the two cases, sharing only 14 proteins in those 45.

Analyzing RMSD measurements across all proteins, the amount of conformational space available to apo proteins is not significantly different than that of holo proteins ($p > 0.05$) (Figure 3-1 and Figure 3-2A, Table 3-1). Most notably, the amount of conformational space *between* apo and holo structures is greater than that *within* either the apo ($p < 0.0001$) or holo ($p < 0.0001$) protein sets (Figure 3-1 and Figure 3-2B-C, Table 3-1). This suggests that the backbones in each of the apo and holo datasets occupy equally sized subsets of the total conformational space available, and there is a great deal of overlap between the two sets. While statistically significant, the difference of 0.86 Å RMSD in apo structures, 0.72 Å RMSD in holo structures, and 1.16 Å RMSD between all structures is less than 0.5 Å RMSD of change. This is likely negligible in the context of an entire protein structure and is close to experimental error, given B-factors for most backbone atoms.

Table 3-1. Averages and Medians of the Maximum Backbone RMSDs.

	Average (Å)	Median (Å)
Apo-Apo Pairs	0.86	0.45
Holo-Holo Pairs	0.72	0.43
Apo-Holo Pairs	1.16	0.69

RMSD values were also calculated specifically for the atoms within the unified binding sites to focus on localized changes incurred upon ligand binding. Binding-site backbone displacement is slightly greater than the whole backbone (Table 3-2). However, the distribution of RMSD by family and type remains largely unchanged (Appendix Figure A-1A-C). These results are observed for both the apo and holo structure subsets (Appendix Figure A-1D-E).

Table 3-2. Averages and Medians of the Maximum Backbone RMSDs for binding site residues only.

	Average (Å)	Median (Å)
Apo-Apo Pairs	1.19	0.31
Holo-Holo Pairs	1.16	0.36
Apo-Holo Pairs	1.80	0.59

Relationships between other metrics have also been investigated. Ligand size would logically impact the magnitude of protein-ligand contact area, and structure resolution can drastically affect our perception of a molecular environment, so it is appropriate to question whether or not these factors have impacted our results. R^2 values between RMSD vs. ligand mass, and RMSD vs. structure resolution were calculated to be < 0.02 at the very best, for all cases. This indicates that no linear relationship is observable between backbone motion and ligand mass, or structure resolution.

3.4.4 Conformational Sampling of Protein Side Chains

Analysis of the protein backbone describes large-scale organizational changes in a protein structure, but it does not answer questions about atomic contacts with ligands. To focus solely on side-chain behavior, we calculated the χ_1 angles for residues within the unified binding sites (see Methods). Comparing χ_1 angles only describe the relative positions of the side chains, not necessarily a degree of flexibility. Therefore, we use the range of χ_1 angles seen across all

structures in a set as a metric for dataset-to-dataset comparisons (see Methods). Comparing χ_1 angle ranges yields information about the extent of occupied conformational space across sets of structures, like all apo structures, all holo structures, or apo and holo structures combined (apo+holo). Distributions of the maximum χ_1 angle ranges are given in Figure 3-3. The maximum χ_1 angle ranges for the apo structures, holo structures, and apo+holo structures are compared for each of the 305 protein families in Figure 3-4. It should be noted that a great majority of the χ_1 ranges increase in all proteins because the apo+holo set has more structures than the apo or holo set alone.

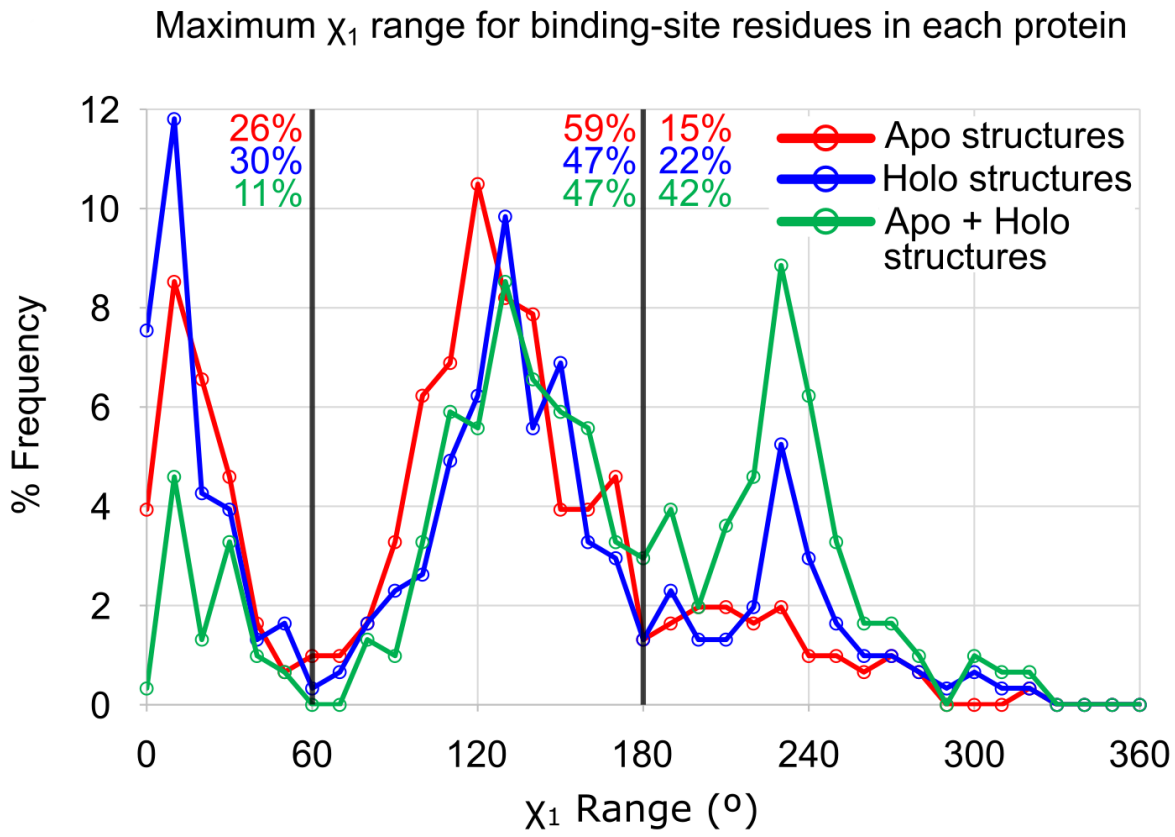


Figure 3-3. Distribution of the maximal χ_1 range in each binding site. Again, the flexibility of the apo and holo states are approximately the same. When the structures are combined, much greater variation is seen in the maximum χ_1 range. The ranges observed across the apo structures are shown in red, and the ranges across the holo structures are shown in blue. The line in green shows the χ_1 ranges measured when the apo and holo structures are analyzed together (apo+holo). The population of structures with maximum χ_1 ranges occupying one conformational well (0-60°), two wells (60°-180°), and all three wells (180°-360°) are given in red, blue, and green numbers for the apo, holo, and apo+holo analysis, respectively.

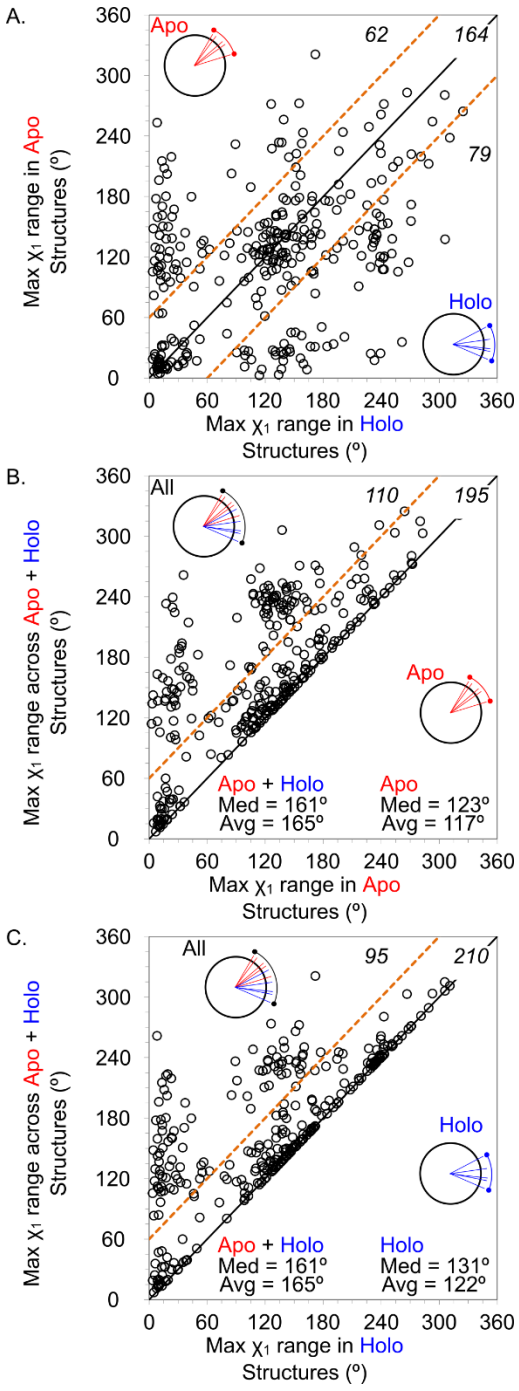


Figure 3-4. Comparisons of the maximal χ_1 range in each binding site. For each protein family, the maximum χ_1 range is given for A) apo vs holo structures, B) apo vs apo+holo structures, and C) holo vs apo+holo structures. The number of points of each section is labeled in black (numbers in parenthesis are the points > 3.5 Å).

The distribution of maximum χ_1 angle ranges shows the most variable side chain for each protein's binding site. The trimodal distribution comes from those side chains occupying one, two, or three of the conformational wells around the χ_1 angle. It is clear that the majority of apo and holo sets have at least one χ_1 angle that spans two conformational wells (ie, the population from 60-180° is largest for apo and holo sets). Only 26% of apo structures, and 30% of holo structures have χ_1 ranges that represent only one energy well ($\leq 60^{\circ}$). When the two sets are combined (the green line for apo+holo in Figure 3-3), there is a significant increase in the number of proteins where the most flexible residue has a χ_1 angle range that spans all three conformational wells available (ie, the population >180°). This shows that in going from the holo to apo state, many systems have side chains pushed into new conformational states not observed in the holo state. This is perhaps better seen in Figure 3-4B,C where roughly one third of the systems show significant displacement of their χ_1 angles (apo+holo χ_1 angle ranges increase by ≥60°).

Traditional statistical tests are not appropriate for the data on maximal χ_1 ranges because the distribution is trimodal. If we examine the average χ_1 ranges for each protein, the data are near-normal in their distribution and appropriate for Wilcoxon signed-rank tests. The average binding-site residue in holo structures exhibits a χ_1 range of 27.5°, while the χ_1 range averages 24.5° in apo structures. The ranges of side-chain motion in holo structures and apo structures are statistically indifferent ($p > 0.05$, Figure 3-5). This follows the trend seen in the RMSD calculations, where the amount of available conformational space to apo structures is approximately the same size as the amount for holo structures. More importantly, the average χ_1 range when combining all structures (holo+apo) is 42.6° ($p < 0.0001$ compared to both apo and holo datasets). This larger range of χ_1 values for all (apo + holo) structures, as opposed to the corresponding apo or holo sets alone, suggests that ligand binding induces rotameric changes in side-chain orientations beyond the threshold of inherent variation.

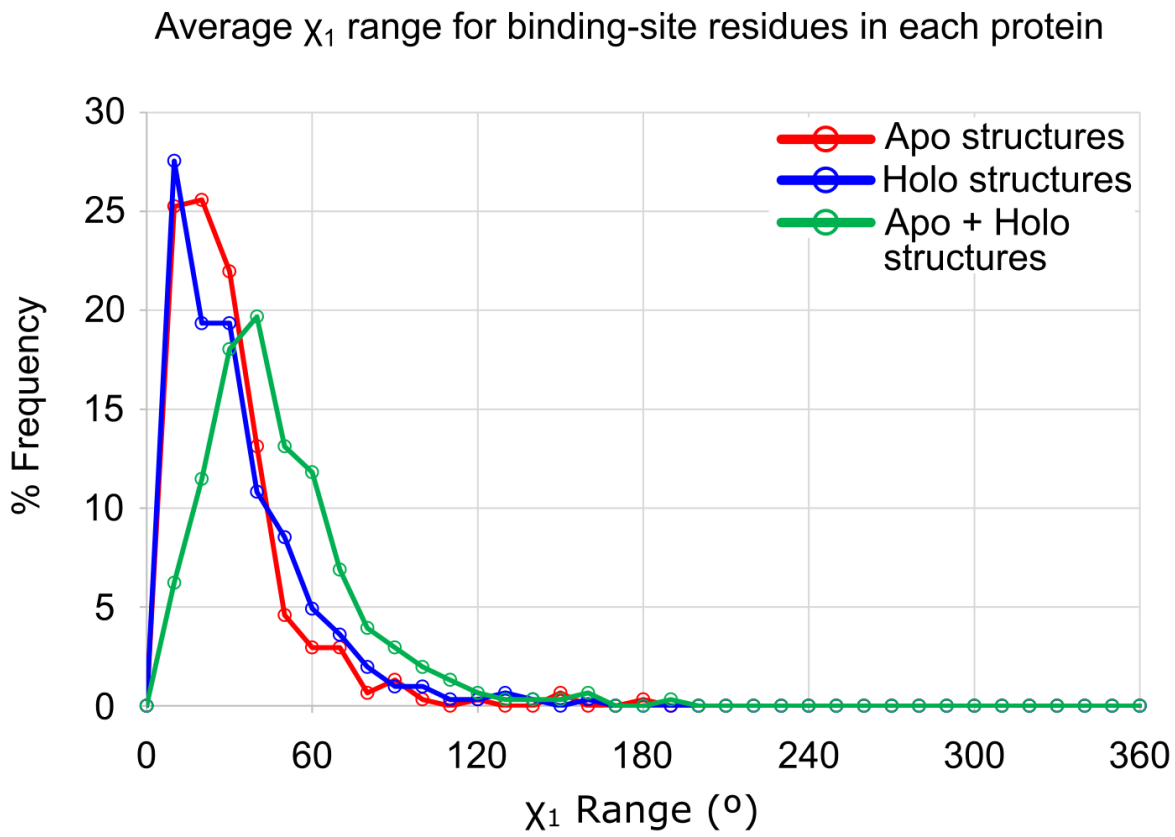


Figure 3-5. Distribution of the average χ_1 range in each binding site. The ranges observed across the apo structures are shown in red, and the ranges across the holo structures are shown in blue. The line in green shows the χ_1 ranges measured when the apo and holo structures are analyzed together (apo+holo). The medians of the average χ_1 range are 19° for the apo structures, 21° for the holo structures, and 37° for the apo+holo structures. The flexibility of the apo and holo states are approximately the same with no statistical significance in their difference ($p > 0.05$). When the structures are combined, much greater variation is seen in the maximum χ_1 range. The difference between the medians of the apo+holo and apo structures is 18° ($p < 0.0001$), and the difference to the holo structures is 16° ($p < 0.0001$).

The difference in side chains between the apo and holo binding sites supports the concept of ligand binding inducing a fit or constraining the binding site.[7] While ligand binding does not generally induce changes of large magnitude the backbone, it has a more significant impact on the side chains. This is supported by a study of B-factors between holo and apo structures where 71% of binding-site protein atoms become less mobile upon ligand binding, and 29% become more mobile.[115]

Again, relationships to ligand size and structure resolution were investigated. All R^2 values for χ_1 range vs. ligand size and χ_1 range vs. structure resolution were < 0.03, indicating that no correlation exists between these factors.

3.4.5 Correlation Between Backbone and Side-chain motion

Correlations between backbone and side-chain motion were assessed by calculation of R^2 values between appropriate datasets using JMP.[113] Comparison of the maximum RMSD vs. maximal χ_1 range for apo-apo pairs, holo-holo pairs, and apo-holo pairs yielded poor R^2 values of 0.02, 0.16 and 0.04, respectively. Lack of correlation between backbone RMSD and χ_1 range suggests that addressing the flexibility of protein backbones and protein side chains is appropriate.

3.4.6 Flexibility of Individual Amino Acids Within the Unified Binding Sites

Establishing that significant changes in side-chain orientation occur upon ligand binding inspired an investigation of the χ_1 angles on a per-amino-acid basis. Radar plots of the occupied χ_1 angles for each amino acid type across the binding sites of all proteins utilized in this study were generated (See Appendix Figure A-2A-R). The χ_1 angles are distributed into three energy wells, with the largest population present where the side chain is gauche only to the N-terminal direction of the backbone. This case is exceedingly prevalent for any amino acids capable of forming an intramolecular hydrogen bond between its side chain and backbone nitrogen. The least common orientation places the side chain gauche to both the N- and C-terminal directions of the backbone, which is a very high energy conformation. Overall, this data shows that side chains in ligand-bound binding sites do not occupy exclusively different conformational space than unbound structures. The larger χ_1 range resulting from calculating with all apo and all holo structures combined simply implies that there are rotameric changes occurring upon ligand binding.

The cumulative distributions in Figure 3-6 display the inherent flexibility of each amino acid type within binding sites of all structures. Important guidelines for incorporating protein flexibility in structure-based drug design may be extracted from the trends in these figures. If residues were allowed to sample 30° of χ_1 conformational space, between 47-90% of side-chain variation could be captured, depending upon the residue type (most flexible Ser and most rigid Trp). In another perspective, trying to capture 90% of all variation would require only about 40° of sampling for the most rigid residue(s), but the most flexible would have to be allowed over 200° of sampling. To represent 90% of the variation in Ser would require 240° of motion, which is approximately the complete range of motion between the three energy wells.

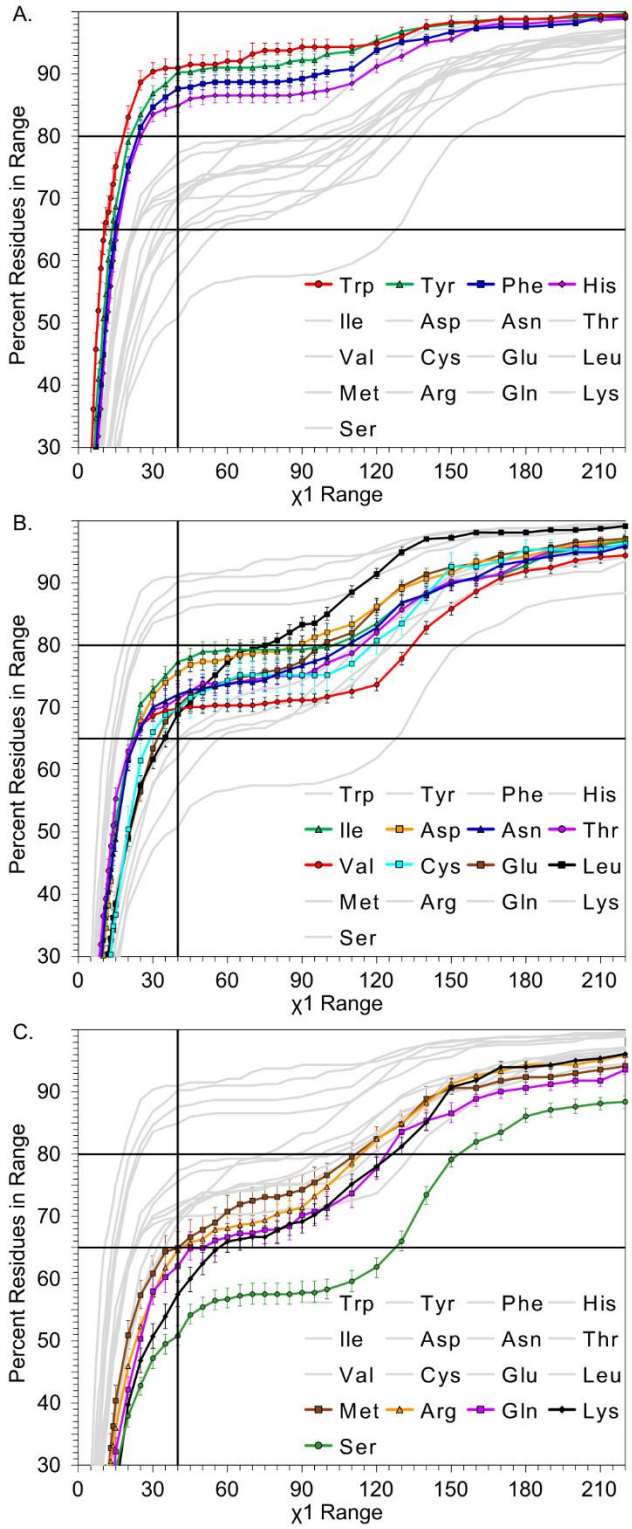


Figure 3-6. Cumulative distributions of binding-site χ_1 ranges for each type of amino acid. The data describes the flexibility of different amino acids as a gradient of rotameric state change. Separated into three groups: A) rigid residues, B) semi-flexible residues, and C) very flexible residues. Error bars represent 95% confidence intervals.

Using this type of breakdown, we rank the amino acids (from least flexible to most flexible) Trp, Tyr, Phe, His < Ile, Asp, Asn, Thr, Glu, Val, Cys, Leu < Met, Arg, Gln, Lys, Ser. Other studies have shown very similar trends with serine and lysine being flexible and the large, bulky amino acids such as tryptophan being rigid, although the ranking is not exact.[105-106, 115] We determine this trend by observing the relative amounts of χ_1 angle range representation at 40°. This trend coincides with classical biochemical intuition, where large hydrophobic residues are more sterically constrained. Not all large polar or charged residues show the same degree of flexibility in χ_1.

This information is immediately applicable in flexible protein docking. Using different thresholds of data inclusion (i.e. what χ_1 range is accommodated with 60% of some residue's data), restrictions could be placed on residues during flexible docking relative to their starting positions (rotameric flip allowed or not). The occupation of observed χ_1 angles can be used to finding "forbidden" rotameric states. Leucine, for instance, almost never occupies the energy well characterized by two gauche interactions with the backbone (2.92% of apo data, 3.41% of holo data in Appendix Figure A-2J).

3.4.7 **Solvent Accessible Surface Area**

SASA calculations have been applied to describe protein-protein binding events [116-117], as well as physicochemical properties of biologically relevant ligands.[118] Figure 3-7 displays the median, minimum, and maximum SASA values for apo structures and holo structures within each of the 305 protein families. There does not appear to be any significant difference observed in SASA for holo structures against their apo counterparts. A distribution of ΔSASA between the minimum SASA apo structure and maximum SASA holo structure for each family is presented in Figure 3-8. The great majority of proteins (72%) have ΔSASA ≤ 100 Å^2, which is rather small. Only 9% of all proteins lose SASA upon binding ligands (ΔSASA < 0 Å^2). These findings agree with Gunasekaran and Nussinov's results suggesting no distinguishable changes in SASA upon ligand binding for flexible, semi flexible, or rigid proteins as classified by other metrics.[99]

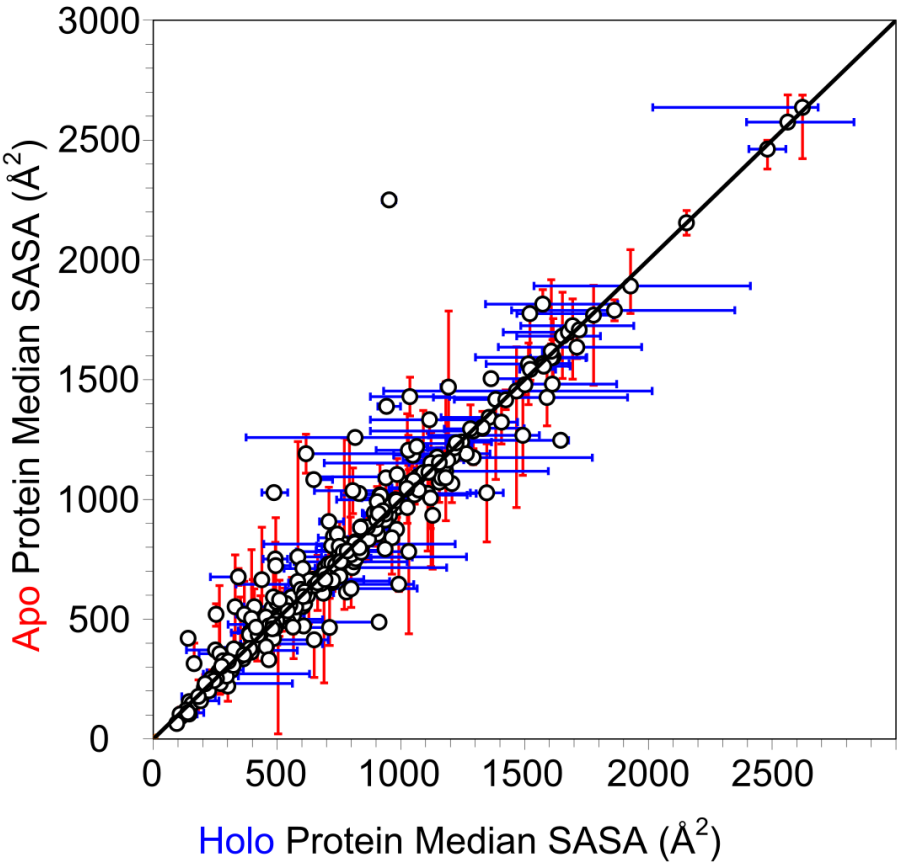


Figure 3-7. Median solvent accessible surface area of unified binding-site residues: apo structures vs. holo structures. Error bars represent the minimum and maximum SASA value in each family for each structure type.

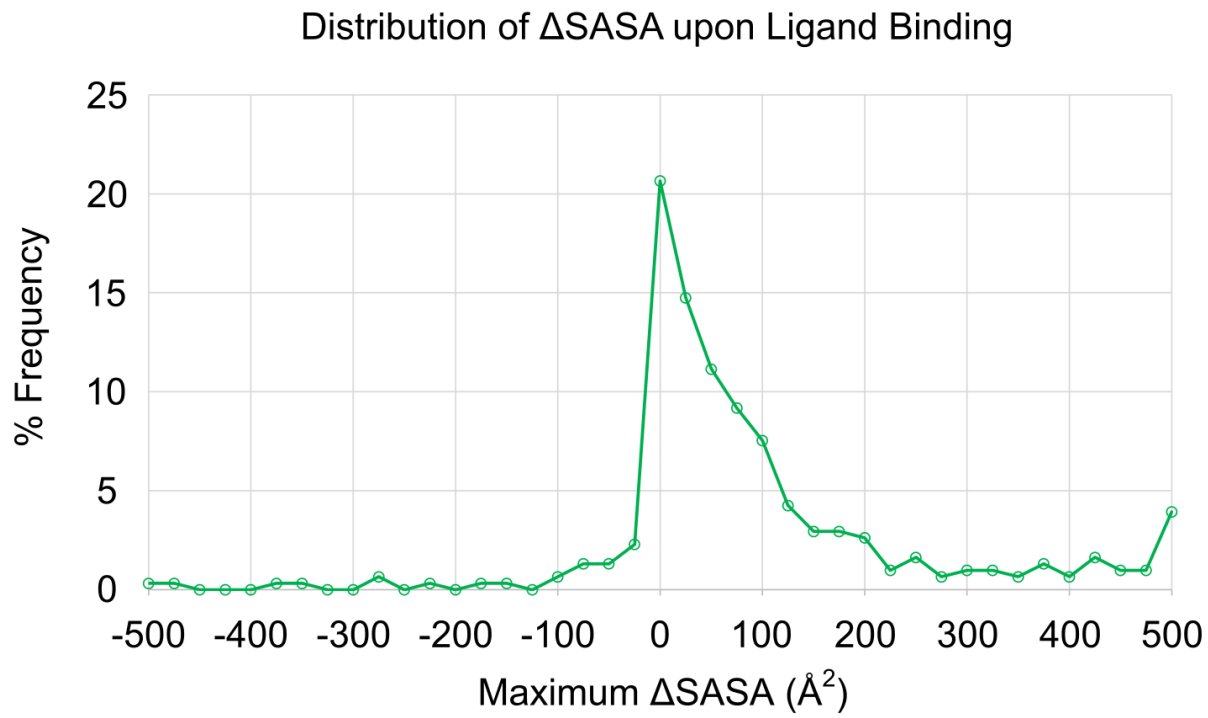


Figure 3-8. Distribution of the maximum change in solvent accessible surface area of unified binding-site residues. ΔSASA was calculated as maximum Holo SASA – minimum Apo SASA.

3.5 Conclusions

Understanding protein flexibility is important in drug design, especially as crystal structures become more widely used as models for binding prediction.[119] 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 backbone and binding site variation among holo or apo structures inherently, as well as what differences arise from ligand binding.

We have shown that ligand-free structures and ligand-bound structures have nearly identical amounts of structural variation, in terms of residual backbone motion (RMSD). A similar range of motion was seen in both the global and binding-site backbones for both the apo and holo structure subsets. The apo-holo pairs showed only slightly larger RMSD.

Examining the side chains through χ_1 angle ranges reveals that apo structures and holo structures have roughly the same flexibility. However, when apo and holo states are combined, the χ_1 angle ranges significantly increase, displaying that binding sites frequently have at least one side chain that gets pushed into new conformations in the presence of ligands.

Through the significant variance in observed side-chain conformations, and relative lack of backbone motion, we support a model of ligand binding where backbone motion is minute, and side-chain flexibility is essential. The lack of correlation between the backbone and side-chain data further suggests that sampling large amounts of conformational space with protein side-chains is not necessarily coupled to having a flexible backbone. Combining these ideas indicates that addressing side-chain flexibility separately from backbone motion is appropriate, which agrees with many modern approaches to flexible ligand docking.

Chapter 4. Binding Site Prediction

4.1 Abstract

Structural biology and genomics projects have provided information about a plethora of proteins with unknown function and interactions. Methodology to identify binding sites is both cost effective and quick, allowing for the characterization of the protein-ligand binding in these new targets. Past studies have focused on reproducing known binding sites based on ligand-bound crystal structures in numerous different datasets. Here, we present a dataset of 304 unique protein sequences (families), represented by 2528 protein structures with at least two ligand-bound and at least two ligand-free structures to represent each family. Unified binding sites representing all bound ligands within each protein family are used for a more robust depiction of the important residues for ligand binding. This work includes a brief survey of six binding-site prediction methods on this dataset: Surfnet, Ghecom, $Ligsite_{csc}$, Fpocket, Depth, and AutoSite. The primary focus of this survey is to examine the performance of ligand-bound vs. ligand-free structures of the same protein. It is commonly believed that it is easier to properly predict binding sites when the pockets are already formed by the presence of a bound ligand. The results reveal that apo structures and holo structures perform equally in the majority of cases. Distributions of the Matthew's correlation coefficients for ligand-bound vs. ligand-free structure performance show no statistically significant difference in structure type vs. performance for Surfnet, Ghecom, $Ligsite_{csc}$, Depth, and Autosite. For Fpocket, there is a statistically significant but low magnitude enhancement in performance for holo structures. The results also show that there is no relationship between 'quality' factors of these crystal structures, such as resolution, and their likelihood of success with the binding-site prediction algorithms. Surprisingly, most families have structures that succeed and others that fail for the same binding site and the same detection method. We expected much higher consistency across varying protein structures of the same sequence.

4.2 Introduction

Interactions between proteins and small molecule ligands are a cornerstone of biochemical function. These interactions vary in specificity, which allows for invention of new molecules to modulate the function of protein targets. Modern drug discovery heavily utilizes structure-based drug design, which requires structural information for the target of interest (typically a protein). As structural information for new targets is obtained, there are cases where little is known about the binding pocket(s) of the protein. Consequently, significant effort has gone into the development of ligand binding-site (LBS) prediction algorithms to help solve this issue. As with many computational methodologies, extensive testing and validation of these algorithms has been a common topic in the literature.[39, 120-121] Unfortunately, due to the timespan of these different validation and benchmarking publications, very few of them use the same dataset. Another common theme among the datasets is the underrepresentation of ligand-free (apo) crystal structures, as most datasets are disproportionately populated with ligand-bound (holo) structures, mirroring the relative population of the Protein Data Bank (PDB)[58].

4.2.1 **Datasets**

The PDB serves as a central repository of structural information for the scientific community. The vast size of the PDB makes it time-consuming and intimidating to harvest data, so the computational community has created resources as focused subsets of the PDB for easier dataset creation, as well as including extra information not available from the PDB. Binding MOAD[3] and PDBbind[61] are web databases focused on protein-ligand interactions, coupling binding data with structural information for drug discovery efforts. ChEMBL[74] and BindingDB[72] are large databases also heavily focused on protein-ligand interactions, but they contain many binding data for which there may not exist structural data. Lastly, there are more specialized datasets. DrugBank[122] is a web database devoted specifically to protein-ligand interactions where the ligands are drug or drug-like molecules, allowing for dataset construction for polypharmacological purposes with ease. LigASite[123] and BioLiP[70] are web-accessible datasets of protein structures both with and without ligands, intended for use in method development of prediction-based computational methodology.

Dataset creation and use are common issues in computational science in general. Previous LBS-prediction algorithms have been trained and tested using numerous different datasets and

databases. The different publications therefore yield vastly different perspectives of how the methods perform. Some methods were trained and tested using manually curated datasets from the PDB[124-126], and some using previously established databases such as LigASite[127] or datasets such as the Astex diverse set.[128] Older publications were more likely to use manually curated datasets, as some of the publicly available datasets and resources were not yet available. Resources such as LigASite[123] and BioLiP[70] have been created with the direct intent for use in prediction method training and testing since that time.

The Astex diverse set is made up of 85 diverse, relevant protein-ligand complexes which is often utilized for benchmarking prediction-based methodology.[129] Although it is a high quality dataset, it is also relatively small and it does not contain any apo structures. LigASite is a dataset of 286 non-redundant proteins with at least one holo and one apo structure for each protein. Though apo protein examples are present in this data set, most proteins are only represented by a single apo structure. Also, the contained holo structures do not all contain biologically relevant molecules. Small molecules such as glycerol, which is commonly used as a crystallographic additive, appear in the top-10 most prevalent ligands in their dataset. The curation process for LigASite did not specifically address common crystallographic additives and inadequately handles small ionic groups for a biological context. LigASite has not been updated since 2012, deeming it less relevant for newer targets. Therefore, we still see a need for more robust datasets for benchmarking computational methodology.

Our dataset is derived from Binding MOAD[3], a collection of high-quality holo crystal structures, followed by sequence-based acquisition of apo structures from the PDB using similar quality criterion for entrance into Binding MOAD. This dataset was culled to 1446 holo structures, and 1082 apo structures, representing 304 unique protein sequences (families) where all structures have resolution of 2.5 Å or better and all holo structures contain ligands that are biologically relevant. The unified binding sites (UBSs) have been calculated for all protein families in this dataset, which represent the union of *all* residues contacted by *any* bound ligand within a family. This dataset is much larger and contains many more ligand-free structures than any of the datasets used in past work cited here. The curation process for this dataset is described in the Methods section. This is a subset of the dataset used in Chapter 1.

4.2.2 **Binding-Site Prediction Methods**

LBS-prediction methods are divided into four categories for discussion: template-based methods (sometimes referred to as genomic-based methods), geometry-based methods, energy-based methods, and other methods.

Template-based methods utilize the atlas of already known protein information as a map to guide the algorithm. Their assumption is that binding sites of new protein sequences may be located using the known binding sites of close structural homologs. Some examples of template-based methods include: 3DLigandSite[130], FINDSITE[131], Firestar[132], I-TASSER[133], IntFOLD[134], and ProBis[80].

Geometry-based methods explore and characterize protein surfaces using a number of biophysical parameters such as Van der Waals radii to locate pockets or clefts. This is because most geometric methods assume that the binding site of a protein is a cleft or pocket in the protein surface. Exploration of the protein surface may be accomplished by calculation of molecular distance, solvent accessible surface area (SASA), and cavity volume. These measurements are computed using probes, spheres, grids, and other forms of spatial voids, which are then clustered or further analyzed to yield ranked cavities presumed to be binding sites. Geometric methods have the advantage of not requiring any prior knowledge about a protein target or any of its close structural relatives, aside from having structural information to work with. This property is advantageous to the purpose of this work. Some examples of geometry-based methods include: SURFNET[124], Ghecom[135], $\text{Ligsite}_{\text{csc}}$[125], Fpocket[136], and Depth[126].

Energy-based methods rely on calculation of phenomena such as hydrogen bonding and pi-stacking to locate regions of the protein where ligands are likely to bind. These LBS-prediction methods are expected to be relatively quick in terms of computation time, so energy-based methods must only account for simple phenomena or make many assumptions to reduce the number of calculations necessary. These methods utilize probe molecules and chemical moieties to generate potentials for locating binding sites. Some examples of energy-based methods include: AutoSite[128], SiteHound[137], Q-SiteFinder[138], and FTSite.[139]

A number of different methods may be categorized as "other" methods, because they either use a different set of physicochemical phenomena not discussed previously or are a combination of different approaches.[39] Some early methods were based on protein sequence, attempting to predict important residues for binding in protein families based on protein sequence or sequence

conservation, but these methods never performed very well.[120, 140] Newer iterations of sequence based methods have been published, but many are designed to target specific types of binding sites (e.g. transition metal and iron-binding complexes[141] or specifically zinc-binding sites[142]). Due to this targeted nature, while they may perform well in their categories, these methods are not robust enough for general ligand binding-site prediction. Even the best performing sequence-based methods are typically outperformed by structure-based methods.[140]

The idea of sequence conservation has been incorporated into other methods, such as the upgrade from Ligsite to $\text{Ligsite}_{\text{csc}}$ by including a re-ranking of top predicted pockets using sequence conservation.[125] Propensity-based methods rank potential binding pockets on a by-atom basis (in the context of likelihood of interacting with a bound ligand) and tally up scores of predicted pockets to either rank novel pockets (e.g. LISE[143-144]) or re-rank pockets from other methods (e.g. STP algorithm[145], Hirayama's method[146]).

Machine learning methods have also been attempted, which utilize any number of previously established physicochemical parameters. These methods utilize computational prediction algorithms ranging from relatively simple Random Forest decision trees all the way to sophisticated neural networks trained on up to dozens of physicochemical parameters. An example is Gutteridge and Thornton's neural network method[147], which predicts the likelihood of a residue being catalytic in nature, where neural network inputs consist of: solvent accessibility, type of secondary structure, depth, cleft the residue resides in, as well as conservation score and residue type. LIBRUS[148] is a support vector machine learner which primarily utilizes sequence-based information, but performed poorly. Similarly, LigandRFs[140] utilizes random forest ensembles to predict binding sites purely from sequence information. Though it was one of the best sequence-based performers, the authors note that structure-based methods still outperform sequence-based ones, even with the help of machine learning.[140]

A last category of "other" methods are ligand-centric methods. These methods develop a specific algorithm or template for locating binding sites of a single molecule of interest (or a small library of molecules). These methods typically utilize machine learning, and could be considered a subset of those methods, but they are unique in their approach of being so targeted to molecules of interest. The UTProt Galaxy pipeline[149] can select and use support vector machines, neural networks, and random forest ensembles to optimize parameters for a user-input molecule and derive a prediction function based on that molecule which will be presented back to the user for

binding-site prediction experiments on targets of their choice. While noteworthy for its unique approach, this methodology is not applicable for the purposes of this work.

There are also newer methods deemed as 'meta-analyses' which combine multiple methods and multiple types of methodology, with some variety of a re-scoring algorithm to try and achieve the best facets of each type of method they use. Some examples include ConCavity[127] which uses 3 methods and MetaPocket[150] which uses 8 methods. While robust, these methods are only available as webservers and are not as user-tunable in order to have seamless integration of the multiple method pipeline. As such, users have limited options to customize parameters for their purposes.

Template-based methods are typically among the best performers during large-scale benchmarking exercises such as the contact prediction section of the Critical Assessment of Protein Structure Prediction (CASP).[151] The LBS-prediction section and its separate assessment in CASP has not appeared since CASP round X.[152-153] The Continuous Automated Model EvaluatiOn (CAMEO) webserver where users may test their server-based automated methods is also of importance for benchmarking.[154] This work is not intended to test the same types of success rate as those large-scale experiments. Rather, the intent is to specifically assess the success rates of ligand-free (apo) crystal structures against their ligand-bound (holo) counterparts for identical protein sequences. Template-based methods have also been excluded from our comparisons, as utilizing libraries of sequence-based template information would inevitably lead to the use of holo structure knowledge to solve the binding site locations in apo structures.

For this work, we elected to use six methods: SURFNET, Ghecom, Ligsite$_{csc}$, Fpocket Depth, and AutoSite. The first five of the methods are geometry-based, while AutoSite is energy-based. Choosing methods was based on two primary factors. First, the availability of source-code to be installed and used on our own machines was required, as using web-servers for large amounts of data was not a viable option. Secondly, we excluded methods that had shown poor performance in previous benchmarks, as detecting performance differences between different structures of the same target is further complicated when the methods are not performing well in general. This list of methods is by no means exhaustive and is simply intended to provide a set of base information as to how LBS-prediction methods perform on different types of structures.

4.2.3 SURFNET

SURFNET is one of the earliest binding site prediction methods, published in 1995.[124] The algorithm grows spheres along the protein surface such that they reach a size where they touch two atoms on their edges, and contain no other atoms. Overlapping spheres as well as spheres with a radius smaller than a user-established threshold are then discarded and the remaining spheres are clustered into cavities. The resulting clustered cavity with the largest volume is assumed to be the putative binding site.

4.2.4 Ghecom

Ghecom (grid-based HECOMi finder) takes spherically-based solvent approaches to a new avenue.[135] The VdW surface of a protein is mapped using solvent probes of varying radii. Utilizing the resulting grid information from the differently sized probes, the algorithm then incorporates mathematical morphology to simplify the protein surface and volume representations by only using small probes to describe areas where larger probes cannot reach. The author of Ghecom proved that this method can reproduce the Connolly volumes derived from other methods.

4.2.5 LIGSITE$_{csc}$

LIGSITE$_{csc}$ is an update of the original LIGSITE method.[125, 155] For the original LIGSITE method, proteins were projected onto a 3D grid. Grid points were considered part of the protein if they were within 3 Å of any protein atom, otherwise they were considered solvent points. Grid points were then scanned for protein-solvent-protein (PSP) events, where vectors in the x, y, and z axis directions are used to connect various series of grid points to find paths that start with a protein grid point, pass through one or more solvent points, and end at a protein grid point. This analysis is done with redundancy, as each grid point may be part of multiple PSP events, which is by design. A threshold is established by the user for how many PSP events are required for pocket detection, and solvent grid points that participate in greater than that threshold number of PSP events are considered part of a pocket. These remaining pocket grid points are then clustered to yield pockets in the protein.

LIGSITE$_{csc}$ provides multiple advancements over the old method. Scanning is increased from 3 axes (x, y, z) to 7 axes by also including the four cubic diagonals. This is only possible because instead of protein coordinates being used to represent the protein, the Connolly surface is used. The method proceeds as follows: First, the protein is projected onto a 1Å 3D grid. The grid

size is minimized by principle component analysis of the protein, followed by a reorientation of the protein on the grid such at the top three determined principle axes are realigned to the x, y, and z directions, respectively. Grid points are then labelled as protein, surface, or solvent. Grid points are part of the protein if there is at least one atom within 1.6 Å. The solvent excluded surface is calculated using the Connolly algorithm.[156] The resulting Connolly surface is a combination of the Van der Waals surface, and the probe sphere surface (with a radius if 1.4 Å) wherever it would touch two or more atoms. Grid points are considered part of the surface if they are within 1 Å of any of the calculated Connolly surface vertices. Due to this close threshold, grid points labelled as surface points will also be labelled as protein points. The algorithm then scans for surface–solvent–surface (SSS) events, utilizing the seven available scanning axes (x, y, z, cubic diagonals). Solvent points which participate in more than the user-specified threshold for number of SSS events are considered to be part of a pocket. The default SSS threshold is 6. Pockets are assembled by a DBSCAN type of clustering using a threshold of 3 Å for adding neighboring points into pocket clusters. At this point, the algorithmic processing would be denoted as $LIGSITE_{cs}$, the final 'c' step to being $LIGSITE_{csc}$ results from re-ranking the top three largest pockets by residue conservation score, taking all residues within 8 Å into account. The conservation score of these sites is derived from the ConSurf-HSSP database.[157]

4.2.6 **Fpocket**

Fpocket is a pocket detection method based on alpha spheres and Voronoi tessellation, which aims to be a freely available counterpart to the Site-Finder algorithm part of the MOE software suite from the Chemical Computing Group.[136] Fpocket first analyzes the atomic coordinates using the Voronoi tessellation package in Qhull. This package returns the Voronoi vertices, atomic neighbors, and vertex neighbors for the coordinates. Alpha spheres, spheres that contact four atoms and do not contain any atoms, are constructed from all possible pairs of Voronoi vertices. Two thresholds, a 'larger' size and 'smaller' size, are established for the alpha spheres, and only spheres above or below one of those thresholds are kept. The resulting alpha spheres are then classified according to the atom types of the four atoms they touch as apolar if at least three of the four atoms are of low negativity (< 2.8, such as a carbon or sulfur in a protein), or polar if they contact 2 or more polar atoms (such as oxygen and nitrogen). The classified set of alpha spheres is then subjected to three types of clustering. First, a rough segmentation pass is accomplished by using the neighbor list output from Qhull which indicates which Voronoi vertices

are connected by a common edge. The algorithm checks if the spheres based on interconnected vertices are close to each other utilizing a distance threshold and clusters them appropriately. Clusters with only one sphere are removed at this point, and the center of mass is calculated for each of the remaining clusters. Next, clusters with closely oriented centers of mass are combined. Finally, a multiple linkage clustering approach is carried out based on the contained vertices in each cluster. If two clusters have enough vertices in proximal space to each other, they are combined. After the clustering is over, the remaining alpha sphere clusters are further pruned by user-set thresholds requiring a minimum number of spheres per cluster, as well as a minimum number of apolar spheres per cluster. The final remaining clusters are then characterized as pockets using a Partial Least Squares fitting and scoring based on implemented pocket descriptors in Fpocket.

4.2.7 **Depth**

Depth is a geometric method based on the relationship between solvent accessible surface area (SASA), and molecular depth.[126, 158] SASA in this method is calculated for the protein target utilizing a rolling-ball method, and normalized against Ala-X-Ala tripeptides to prevent over-attribution of accessibility.[159-160] Molecular depth is defined as the distance between a molecule (as an average of the distances of all of its constituent atoms) and bulk solvent. In the case of binding site prediction, this is the depth of amino acid residues in the protein sequence from the bulk solvent outside of the globular protein.

The solvation process begins with placing a protein in a pre-equilibrated box of SPC216 water such that all residues are buried by at least two hydration shells (5.6 Å). All clashing waters, those within 2.6 Å of any protein atom, are then removed. Non-bulk water clusters are then removed *via* calculations involving water neighborhoods (waters contained within a spherical distance from the atom in question). The user may set two variables here, the number of minimum waters required in the water neighborhood, and the radius size that dictates this neighborhood. The default values imply water neighborhoods containing at least 4 other water molecules within a 4.2 Å distance, which the authors assess as being 1.5 hydration shells.

This process is iterated until there is no further removal of water from the solvent box, implying that all non-bulk water has been removed. This entire process is repeated to achieve realistic implementation of solvent diffusion. For each pass of solvation, the protein inside the water box is rotated at a random angle along a random axis that passes through the protein's center

of mass. Following this random rotation, the protein is then shifted an arbitrary distance (<2.8 Å, the distance between any two waters in the grid) along the X-axis. The user may set how many times this re-solvation process is performed, but the default value is 25 solvations.

Binding cavities are then determined under the assumption that the only remaining water is bulk water outside of the protein surface, and a pocket would otherwise have some solvent in it, therefore having higher SASA. Pocket residues, therefore, have large molecular depths, and higher SASA. The algorithm then assigns probability values based on SASA and residue depth to each amino acid in the protein sequence. Those residues with probabilities above a user-alterable threshold are then kept as potential binding cavity residues. The protein is then re-solvated (25 times by default), using the same process as above, except that only clashing waters are removed this time. Water molecules within 4.2 Å of any potential binding cavity residue are noted, and any water within 4.2 Å of those retained waters are also noted. These noted water networks are then used to determine the rest of the binding site residues that are then assembled into the predicted cavity provided to the user.

4.2.8 **AutoSite**

AutoSite is an energy-based method which uses AutoDock[161] affinity maps computed with AutoGrid4[162] for three generic atom-type grids to identify binding sites. The three AutoDock generic atom types are hydrophobic (carbon, C), hydrogen-bond acceptor (oxygen, OA), and hydrogen-bond donor (hydrogen, HD). These maps are regularly spaced (1.0 Å, in most cases) grids made up of one of the three generic atom types. The computed affinity maps yield information about the sum of all interaction energies between each grid point and all receptor atoms in its local area. The affinity maps only represent atom-specific affinities and do not include electrostatics or charge-based desolvation. AutoSite computes probe maps which cover the entire protein surface and then selects high affinity points from each map based on affinity cutoffs, which are probe-specific. The algorithm then merges the three sets (for each atom type) of high affinity points into a composite map by selecting the minimum value at each grid position. The resulting points are then clustered to find putative binding sites.

4.3 Methods

4.3.1 Dataset Construction

Holo structures were derived from Binding MOAD[3], a source of high quality protein-ligand complexes that have a maximum of 2.5 Å resolution. Biologically relevant ligands are differentiated from opportunistic binders (e.g. salts, buffers, phosphate ions) in the crystal structures of Binding MOAD, making curation of relevant ligand structures straightforward. Furthermore, use of Binding MOAD excludes covalently bound ligands. Structures with more than one valid ligand were excluded from this study in favor of binary protein-ligand complexes to ensure that only one pocket was being analyzed in each protein. Any structures containing additional molecules in their binding site, such as additives, were also excluded.

Holo structures were then clustered by 100% sequence identity in both directions, without replacement (to ensure a non-redundant dataset). A subsequent 95% sequence identity clustering of those families was then performed to suggest any families that should be merged due to simple N or C terminal amino acid additions. Sequence identity between structures was determined using BLAST.[85] Any families differing in protein core sequence were kept separate.

Apo structures were then cultivated from the PDB using the same bidirectional 100% sequence identity BLAST procedure, requiring better than 2.5 Å resolution.[59] Structures were screened for bound molecules, and only those containing acceptable additives or no additives at all were kept. Acceptable additives were restricted to HET groups of 5 atoms or less and a MW of 100 Daltons or less. Each HET group was inspected for chemical appropriateness.

Finally, proteins that did not have at least two holo proteins and two apo proteins were excluded from the dataset at this point.

4.3.2 Family Size Reduction

After construction of the UBS (described below), but before binding-site prediction, protein families with more than 10 structures of a single type (apo, holo) were reduced to 10 of those type of structures utilizing the following procedure: Exhaustive pairwise RMSDs were calculated for each family (every possible apo-apo, apo-holo, and holo-holo combination) using Gaussian weighted RMSD methodology developed previously in our laboratory.[111] Matrices were constructed for holo-holo pairs, and separately apo-apo pairs, for families in need of reduction. These matrices were then utilized in PAM clustering (Partitioning Around Medoids) in the R

statistical package to determine the 10 most diverse structures to represent a family at hand.[114] For example, the largest family (family 1) of Lysozyme had to be reduced from 280 apo structures to 10 apo structures (utilizing a 280x280 pairwise RMSD matrix).

Theoretically, because the entirety of the holo structure set is used to construct the UBS prior to this data reduction, their influence on the outcome of the experiment remains. This reduction was only intended to reduce computation time for the prediction methods. These methods reduced the dataset from 2369 holo and 1679 apo structures, to 1448 holo and 1026 apo structures. Lastly, due to poor binding site resolution in structure 1HNK, which resulted in having 10 binding site residues unresolved, its entire family (two apo, two holo) was removed from the dataset. This results in the final dataset of **304** protein families, with **1446** holo and **1082** apo structures.

4.3.3 File Choice, Setup, and Preparation

These steps were taken prior to any binding-site calculations. The first biounit model containing the relevant ligand of the corresponding PDB structure was used by default for each structure. All hydrogens were removed from the files. All ligands and waters were removed from the files.

All protein systems were renumbered utilizing the pdbSWS database prior to binding site calculation and assembly.[82] In the cases where this would result in more than one numbering pattern inside of a family, one structure's numbering was applied to the other structures. If this was not possible and there was no method to renumber a structure to the same pattern as the rest of its family apart from manual processing, it was discarded from the dataset out of consideration for reproducibility.

Renumbering structures was necessary because some structures were numbered differently (especially common when going between apo and holo structures). Protein numbering becomes critically important in the case of UBSs, where it is necessary to harvest residue data from the UBS when there are no ligands present to define the site (apo structures).

After binding sites were identified (detailed in the following section), the files were reduced to contain only the chain(s) involved with a single copy of a binding site.

4.3.4 Binding Site Identification and Compilation of the "Union" Binding Site (UBS)

The binding site was defined to include all protein residues within 4.5 Å of any biologically relevant ligand for each protein, which should capture both hydrogen-bonding and van der Waals interactions. Hydrogen atoms were not considered during this 4.5 Å calculation (for either the protein or the ligand). Most of the crystal structures for a given protein had different ligands bound, so many could have a slightly different set of residues near the ligand. Therefore, the summation of all sets of residues in all complexes for each protein was used to identify the "union" binding pocket for that protein, i.e., unified binding site (UBS).

4.3.5 Responding to Computational Errors

Structures which resulted in errors when submitted to a particular method were very uncommon, and most of the time reformatting the PDB file in some manner alleviated the issues (eg. the multiple residue conformation issue detailed in the AutoSite section 4.3.8.6). Structures which produced errors for the various methods are provided below in Table 4-1. Notably, none of these structures were problematic with more than one method.

Table 4-1: PDBids for structures which resulted in system errors for the various LBS-prediction methods. Apo structures are denoted in red, holo structures are denoted in blue.

Surfnet	Ghecom	Ligsite	Fpocket	Depth	AutoSite
3n5k	4ey1	3o4g	None	None	1a16
1su4	1g7b	1ve6			1e3z
	1tym	2hu7			1h2j
	4ajz	2ogz			1e43
		2hu5			1tgb
					2vb9

Importantly, failures of 3n5k and 1su4 with Surfnet occur due to the algorithm attempting to generate an interaction array which is larger than a hard-coded threshold value. We opted to not edit the source code to fix this error.

4.3.6 Responding to Empty Prediction Files

For all but one method (Fpocket), structures yielding no predicted pockets was an extremely rare occurrence (Table 4-2). Fpocket yielded no predicted pockets for 40 different structures (17 apo, 23 holo). Yielding zero pockets resulted in a score of zero for precision, recall,

F score, and MCC. These failures were double checked by-hand as single command-line submissions, to ensure no other issues were taking place.

Table 4-2. PDBids for structures which resulted in no predicted pockets for the various LBS-prediction methods. Apo structures are denoted in red, holo structures are denoted in blue.

Surfnet	Ghecom	Ligsite	Fpocket Apo	Fpocket Holo	Depth	AutoSite
N/A	1g7b	1ve6	1aki	1a7x	2olz	1b2d
	4ey1	3o4g	1b2d	1b0d		1n40
	1tym	2hu5	1g7b	1j4h		1vie
	4ajz	2hu7	1guj	1our		2vjz
		2ogz	1mi7	1tym		1uof
			1rnu	1uzv		1vif
			1u1t	1zt9		2oly
			1uoj	2boj		2rk2
			1yy6	2oly		3lb2
			1zz6	2olz		4ajz
			2rh2	2z3h		
			2vjz	3dcq		
			3a93	3ipe		
			3az5	3qe8		
			3w3b	4ajx		
			4bwo	4ajz		
			4f4t	4b4q		
				4b4r		
				4joj		
				4jor		
				4lkd		
				4tun		
				4tz8		

4.3.7 **Assessment Metrics**

There are numerous metrics for assessing predicted pockets. In the publications presenting these LBS-prediction methods, assessment metrics were often specific to the method being presented, and some proximity calculations to actual binding site residues or ligand atoms were also used. For instance, Fpocket's metrics (PocketPicker criterion and Mutual Overlap criterion) were based around the relationship between the alpha spheres in the predicted pocket and the atoms of the bound ligand.[136] For the purpose of this work, methodology that required reference to the ligand atoms was avoided, as our binding sites are defined using many ligands. Metrics that utilize specific aspects of some methodology (such as those using alpha spheres) was also avoided, as the

assessment metrics need to be applicable to all methods used. Receiver Operator Characteristic (ROC) curves are a classic method for analysis of these types of data. However, it has been thoroughly discussed that analyzing ROC curves for performance of functional residue prediction can be highly misleading.[148, 163]

We have therefore chosen to assess methods utilizing metrics intended for binary classification events revolving around the four elements of the resulting confusion matrix: **T**rue **P**ositives, **F**alse **P**ositives, **T**rue **N**egatives, and **F**alse **N**egatives. Both precision/recall analysis, as well as Matthew's Correlation Coefficients (MCCs) have proven to be useful in the assessment of prediction methods.[127, 135, 158] As such, we use MCCs, as well as F scores, which are calculated from precision and recall, as metrics to represent the predictive power of the various methods. The calculation of these metrics is presented below in Scheme 1.

Scheme 1: Formulae for Matthew's Correlation Coefficient (MCC), Precision (P), Recall (R), and F score (F).

$$TP = True\ Positive, FP = False\ Positive, TN = True\ Negative, FN = False\ Negative$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F = 2 * \frac{P * R}{P + R}$$

Precision is value between 0-1 where 1 is a perfect score, representing the likelihood of a method's predictions to be correct. Recall is a value between 0-1 where 1 is a perfect score, representing what percentage of the true correct answer is represented by the true positives predicted by the algorithm. The F score is a value that represents the harmonic mean of precision and recall. In our analysis, F score values were presented instead of P or R values because they simplified the predictive power of a method into a single number for easier comparison between different methods or data types (holo vs. apo).

Precision and recall do not account for true negatives in any way and are thus blind to the relative ratio of possible answers that could be derived; in this case, this alludes to the size of the

binding site in relation to the size of the protein. While going as far as rewarding methods for correct true negatives (metrics such as accuracy) would be disadvantageous to the purpose of this work, MCCs are a good medium where correct true negatives are not rewarded, and false positives are still punished albeit less severely. The MCC also has the advantage of being one stand-alone metric, where precision and recall need to be condensed to an F score to provide a single figure. Both P/R and MCCs have been used in LBS-prediction benchmarks in the past. We will present both F scores (representing P and R) as well as MCCs for a more robust analysis.

4.3.8 Prediction Method Parameters

4.3.8.1 Depth

Depth[126, 158] was run with these conditions set in the parameters file: detection threshold of 0.8, cavity size of 4.2, resolvation cycles set to 5, solvent shell size of 4.2 Å, 25 depth cycles, minimum number of required solvent neighbors set to 4, ASA resolution of 92, ASA probe radius of 1.4 Å, and USE_MSA set to 1. A web version of DEPTH, as well as a download mirror for the software can be found at: http://cospi.iiserpune.ac.in/depth/htdocs/intro.html

4.3.8.2 Fpocket

Fpocket[136] settings were left at their default values for pocket detection. The defaults are: minimum alpha-sphere radius (3 Å), maximum alpha-sphere radius (6 Å), minimum apolar neighbors for apolar consideration (3), minimum a-sphere per pocket (30), maximum first cluster distance (1.73 Å), maximum distance for single linkage clustering (2.5 Å), minimum number of neighbors close to each other (3), maximum distance between two pockets' barycenter (4.5 Å), minimum proportion of apolar spheres in a pocket (0), number of Monte-Carlo iterations for the calculation of pocket volume (2500). Information about Fpocket, as well as a download mirror, can be found at: http://fpocket.sourceforge.net/

4.3.8.3 Ghecom

Ghecom[135] was run with large probes (mode = 'P'), and the top binding site was defined as cluster #1 in the output clustered PDB file. The web version of Ghecom, as well as a download mirror for the software can be found at: http://strcomp.protein.osaka-u.ac.jp/ghecom/

4.3.8.4 LIGSITE$_{csc}$

LIGSITE$_{csc}$[125] was run with the number of pockets set to 1 (-n 1), using '-i' to direct to input files through a wrapping script, with the rest of the parameters set to their default values (1 Å grid space, SSS event threshold set to 6, surface density 0.5). As pockets are provided as a centroid atom of the surface cluster, residues within an 8 Å sphere were back-calculated to represent the binding pocket. This protocol is derived from that of the authors.[125] A web version of LIGSITE$_{csc}$ and a download mirror can be found at: http://projects.biotec.tu-dresden.de/pocket/

4.3.8.5 SURFNET

SURFNET[124, 164] parameters are as follows: gap files were generated with the following parameters (N, N, Y, 4.5) for: **N**one map-format, SITE records **N**ot required for mask region, and **Y**es to requiring neighboring atoms for mask region with a **4.5** Å cutoff (the same cutoff used for defining the binding sites from the original ligands). Binding site residues were then extracted from the generated gap files for each structure. SURFNET's web portal can be found here: https://www.ebi.ac.uk/thornton-srv/software/SURFNET/

4.3.8.6 AutoSite

AutoSite[128] requires PDBQT format files, which are a proprietary file format for the AutoDock suite of tools. Before generating PDBQT files for the dataset, scripts were run to remove any multiple-occupancy resides from the initial biounit files (eg. ASER, BSER, where the two occupancies would sum to 1). The highest occupancy representation for each residue was kept. This process was necessary because the PDBQT file conversion process does not accommodate multiple occupancy residues well, and results in ATOM section lines with >80 characters that are unreadable by any PDB parser.

The PDBQT files were then generated using Autodock Tools. AutoSite was run with default settings, and the top predicted binding site cluster was analyzed for each protein structure (XXXX_cl_1.pdb). Actual predicted binding site residues were back-calculated from these point clusters using a 4.5Å distance cutoff. AutoSite's web portal can be found here: http://adfr.scripps.edu/AutoDockFR/downloads.html A guide for preparing PDBQT files can be found here: http://autodock.scripps.edu/faqs-help/how-to/how-to-prepare-a-receptor-file-for-autodock4

4.4 Results and Discussion

4.4.1 Dataset Properties

The most recent release of Binding MOAD was clustered using a very strict sequence identity cutoff to obtain relevant holo structures, and matching apo structures were obtained from the PDB as described in the methods section.[3] Upon filtering for proteins with at least 2 holo structures and 2 apo structures and reducing all families to a maximum of 10 structures for each apo/holo state (see methods), this dataset reduces to 304 different proteins, represented by 1446 holo structures and 1082 apo structures.

The proteins with the most holo structures prior to family size reduction are carbonic anhydrase II followed by trypsin, with 174 and 120 holo structures, respectively. The proteins with the most apo structures before family size reduction are lysozyme followed by ribonuclease-A, which have 280 and 79 apo structures, respectively. This redundancy is accounted for in two major ways. First, when describing prediction assessment for each protein sequence (family), the value will be given as an average, median, maximum, or minimum for the entire family as one value to represent all contained structures. Second, families with more than 10 of either type of structure are reduced to the 10 most diverse (*via* RMSD) representatives for prediction calculations. For example, the carbonic anhydrase II family has 174 holo structures, and all of the ligands for the 174 structures are used to build the UBS, so all structures are truly represented; however, only the 10 most diverse holo structures are used in the prediction calculations to save computational time. This process is detailed in the Methods section. The results of this family size reduction are 2528 protein structures (1446 holo, 1082 apo) which are actually tested with every one of the six LBS-prediction methods.

The biologically relevant ligands that occupy the holo structures in this dataset are diverse and represent many different classes of molecules. The average molecular weight (MW) of the ligands is 374 g/mol with 80% of ligands less than 500 g/mol and 95% less than 800 g/mol. This large range in molecular size helps with building diverse UBSs. The distribution of UBS sizes and number of each residue type represented across all binding sites are presented in Figure 4-1A-B.

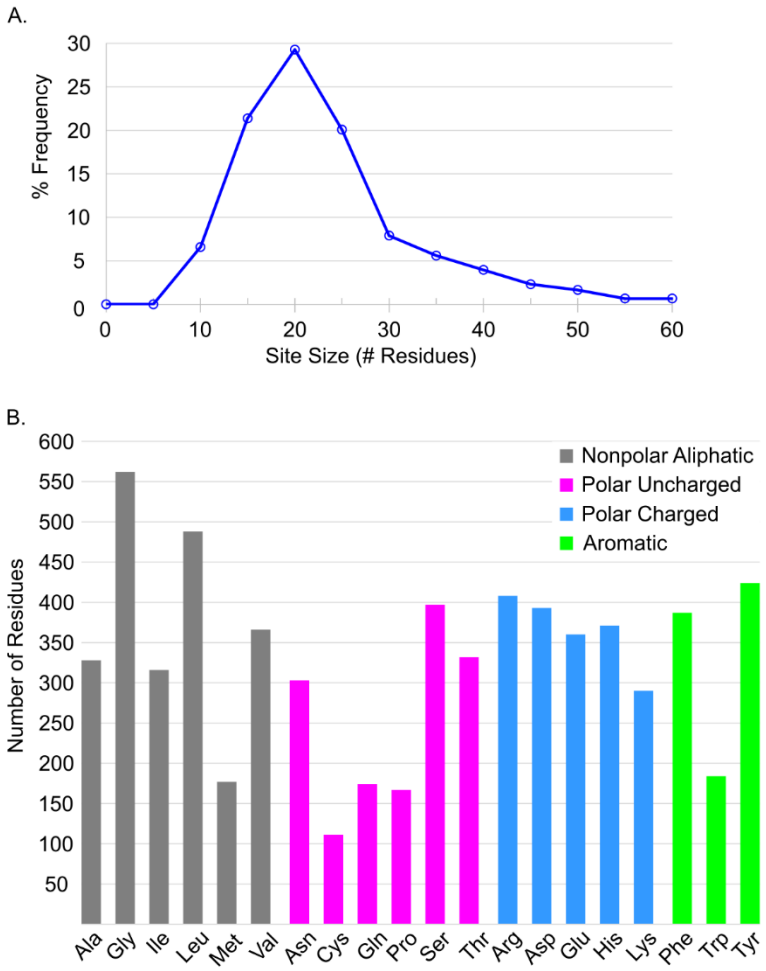


Figure 4-1. A) Distribution of the sizes of unified binding sites for the 304 protein families in this dataset, as % frequency. B) Distribution of amino acid composition of the 304 unified binding sites.

4.4.2 LBS Prediction

Predictive power is assessed using two metrics in this work: F scores and Matthew's Correlation Coefficients (MCCs). Justification for this type of analysis and description thereof can be found in Methods. Comparison between performance of different methods, or of different sets of data (apo vs. holo), will be represented by p values from Wilcoxon rank-sum tests.

Biounit files for all 2528 protein structures were prepared as described in the methods section. All structures were submitted to each of the six LBS-prediction methods and the top predicted pocket of each method was analyzed. For the methods that yield a grid representation of the binding site rather than actual binding site members (AutoSite, Ligsite_{csc}), the binding sites were back-calculated using a distance cutoff of 4.5 Å unless a different cutoff was specified in the

citation for the method (8 Å for Ligsite$_{csc}$[125]). Any structures which did not yield any predicted pockets were assigned zero values for MCC, precision (P), recall (R, also referred to as sensitivity or true-positive rate), and F score, after they were inspected to ensure the programs completed their calculations properly. The procedure for dealing with structures that resulted in errors for the various methods, as well as a list of these very few structures, is provided in the methods section. Analysis metrics were then calculated for the rest of the resulting structures using in-house parsing scripts.

Our analysis of the predictive power for the six LBS-prediction methods begins by presenting distributions of F scores for all methods (Table 4-3, Figure 4-2A-F). These are distributions of the family medians, divided into the subcategories of apo structures and holo structures. Importantly, Wilcoxon rank-sum analysis of apo vs. holo distributions yields $p > 0.05$ for all methods except for Fpocket ($p = 0.04$). On first inspection of Figure 4-2, it may appear as though a large percentage of structures fail to run in Ghecom, Ligsite$_{csc}$, Fpocket, and AutoSite, but data points falling in the zero-score area are structures that either do not predict the correct binding site as their #1 predicted site or structures where no site is predicted at all (a rare occurrence, see methods section 4.3.6).

Table 4-3. Median of family median F scores and MCCs for apo and holo datasets for all six LBS-p methods. Wilcoxon p values are the same as those found in Figure 4-2 and Figure 4-3.

	Apo F	Holo F	Wilcoxon *p*: F score	Apo MCC	Holo MCC	Wilcoxon *p*: MCC
Surfnet	0.23	0.23	> 0.05	0.22	0.23	> 0.05
Ghecom	0.48	0.54	> 0.05	0.50	0.53	> 0.05
Ligsite$_{csc}$	0.49	0.52	> 0.05	0.47	0.50	> 0.05
Fpocket	0.42	0.53	0.04	0.43	0.52	0.03
Depth	0.40	0.42	> 0.05	0.38	0.40	> 0.05
Autosite	0.32	0.46	> 0.05	0.28	0.43	> 0.05

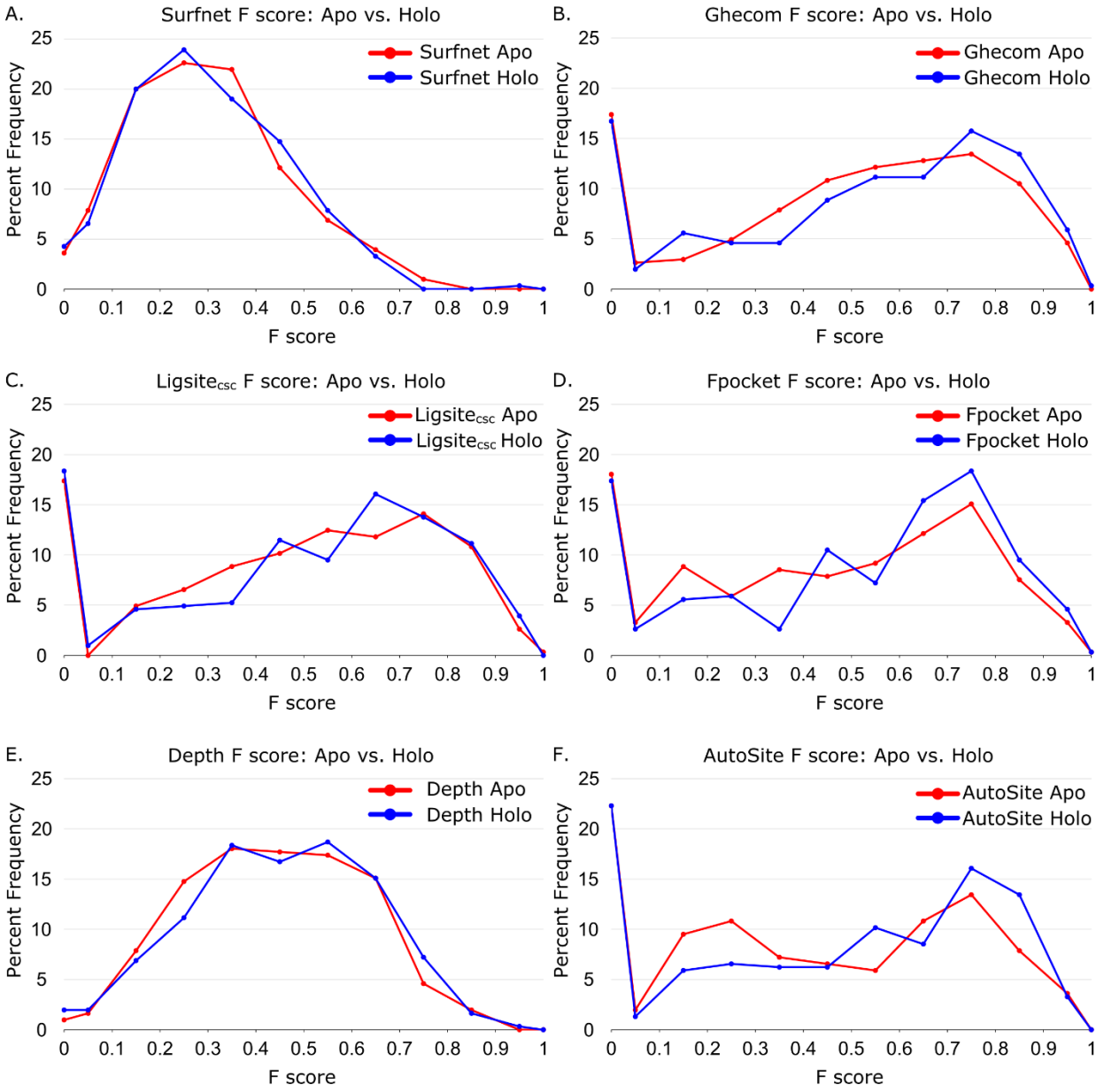


Figure 4-2. Distribution of family median F scores of apo and holo protein structures. Data presented for A) Surfnet ($p > 0.05$), B) Ghecom ($p > 0.05$), C) Ligsite$_{csc}$ ($p > 0.05$), D) Fpocket ($p = 0.04$), E) Depth ($p > 0.05$), and F) AutoSite ($p > 0.05$).

While Fpocket does appear to have a slight performance preference for holo protein structures, the magnitude indicated by the *p* value is not large. This implies that the predictive power for these methods is not heavily impacted by the presence of a pre-organized binding site with a ligand in the starting structure. The same trend is observed when using MCCs as the evaluation metric of predictive power (Table 4-3, Figure 4-3A-F). Only Fpocket ($p = 0.03$) has a

statistically significant ($p < 0.05$) correlation between predictive power and structure type (holo vs. apo), again suggesting that holo structures perform slightly better with this method, but the trend is weak.

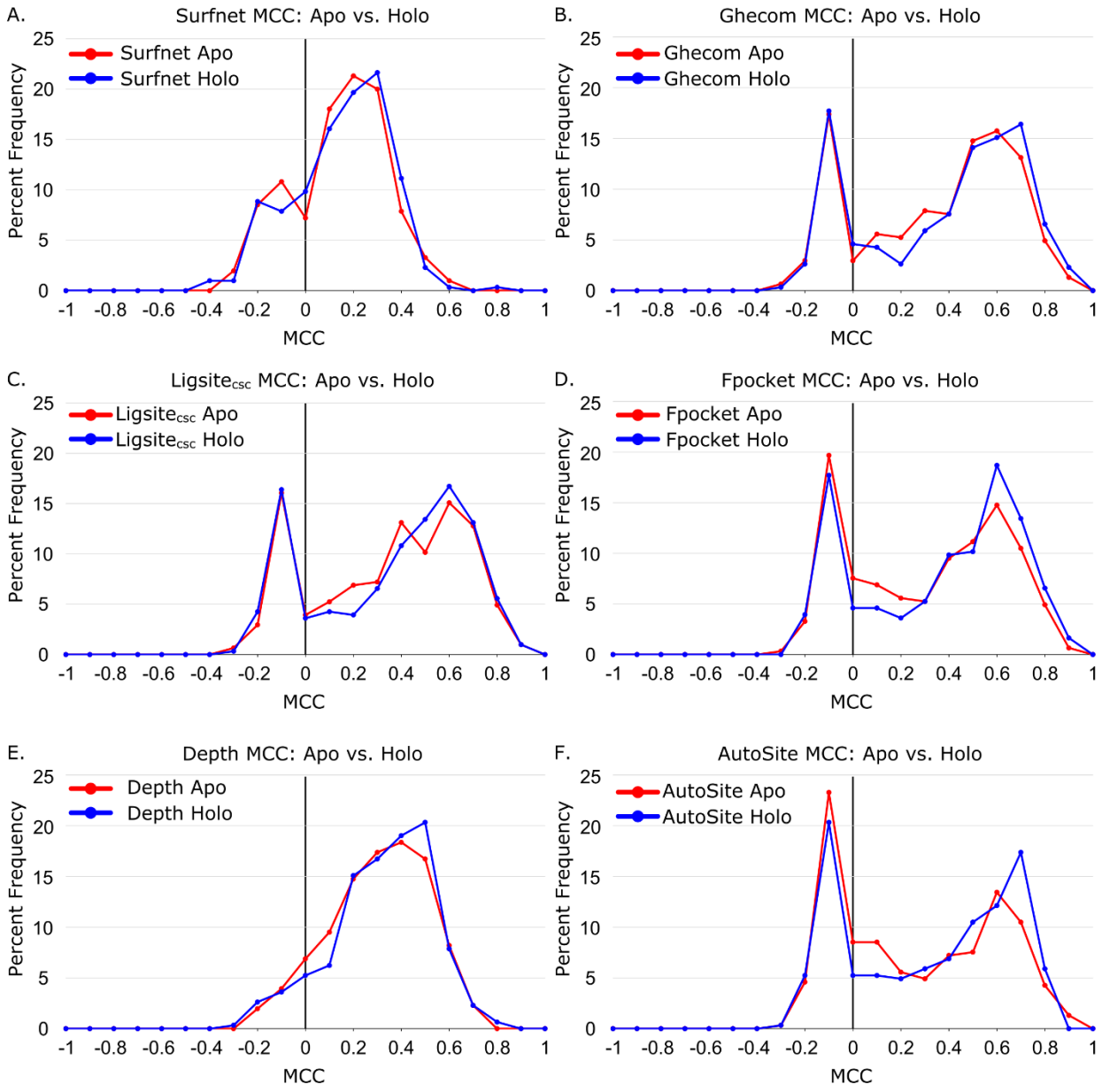


Figure 4-3. Distribution of family median Matthews Correlation Coefficients (MCCs) of apo and holo protein structures. Data presented for A) Surfnet ($p > 0.05$), B) Ghecom ($p > 0.05$), C) Ligsite$_{csc}$ ($p > 0.05$), D) Fpocket ($p = 0.03$), E) Depth ($p > 0.05$), and F) AutoSite ($p > 0.05$)

For both MCC analysis and F scores, two primary patterns of predictive power are observed. Surfnet and Depth appear to have a higher likelihood of mediocre predictive power (F score < 0.7, MCC < 0.6), while also having a much lower rate of complete failure (scores near zero). The other four methods appear to have a more bimodal distribution of scores, either accurately predicting a relatively large portion of the binding site or failing completely in their top predicted site (18-22% of protein families).

Of the 2528 protein structures (1446 holo, 1082 apo) processed with these methods, only five structures failed (zero correct binding site residue predictions, R = 0) in every one of the six methods. Additionally, 1053 structures failed to predict any part of the binding site (R = 0) in at least one method, but 952 of those structures have at least 50% of their binding site predicted by at least one other method (R > 0.5). The performance of many structures appears to be dissimilar between the methods. Exhaustive comparison of the resulting F scores and MCCs for each individual PDB structure between every combination of the six LBS-prediction methods was performed, resulting in $R^2 < 0.1$ for every comparison. This shows that the performance of any given structure with one method provides no indication of how that structure will perform with another method.

Another analysis for the success of each method is to view the F-scores and MCCs as a by-family comparison between the two structure types (i.e. how do the apo structures of a given protein perform relative to the holo structures of the exact same protein?). Using family medians for the representative family data points, and family minima/maxima as error bars, the predictive power of the six methods is presented for the F scores in Figure 4-4A-F and MCCs in Figure 4-5A-F. Interestingly, family maxima and minima span the gamut of performance for each method for nearly all of the 304 protein families in both F scores and MCCs. *For most of the protein sequences (families) in this study, there simultaneously exist high-quality structures which a given method can accurately predict the majority of the ligand binding site, as well as structures where the same method completely fails to identify the same binding site as the top site.* This is true for both the apo and holo states of the proteins and has serious implications for benchmarking LBS-prediction methods, as the choice of protein structure greatly influences outcome. This inherent variability makes it impossible to rank methods and points to a need for greater consistency on the part of the methods.

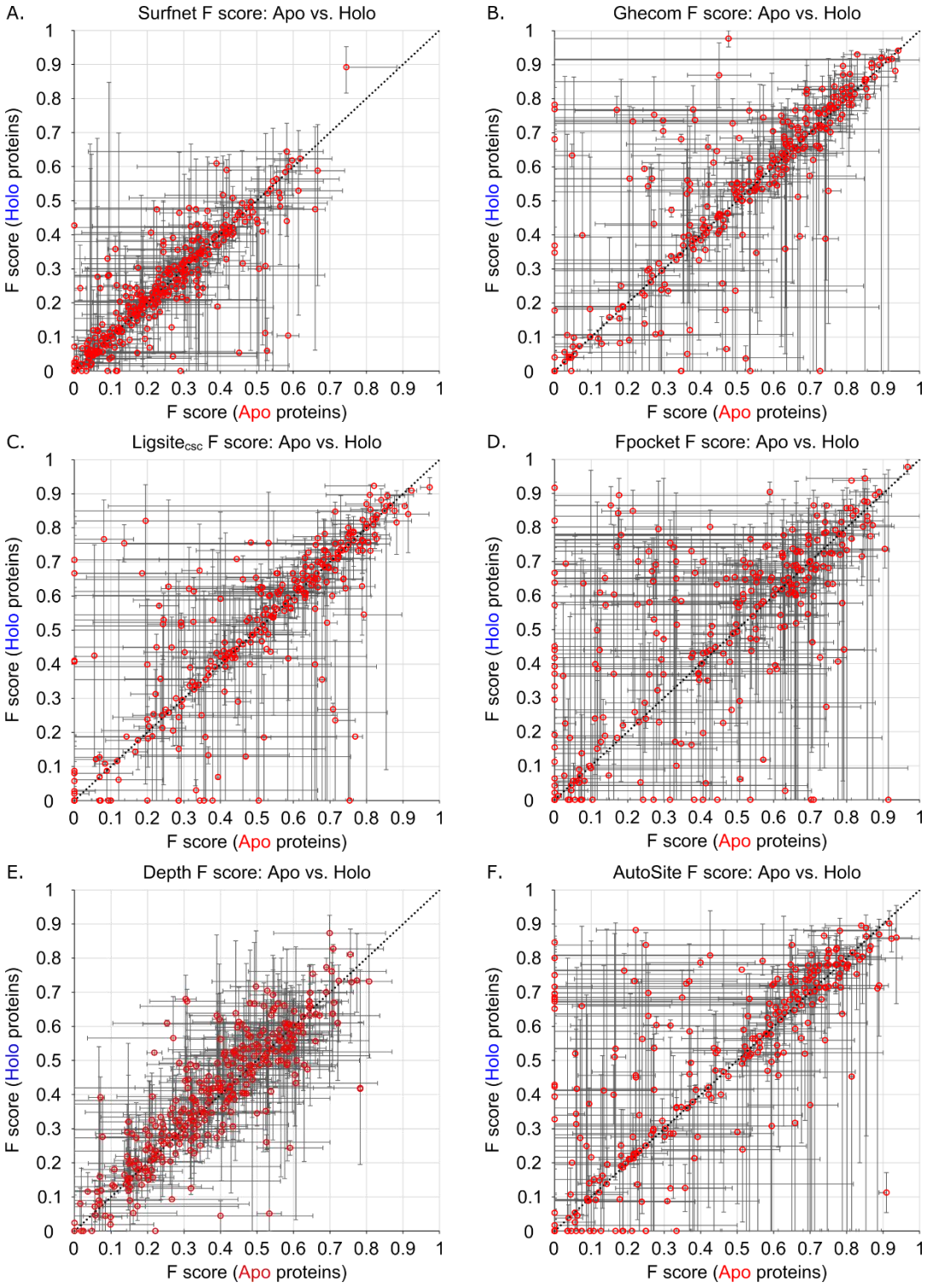


Figure 4-4. Family median F scores of apo and holo protein structures. Data presented for a) Surfnet, b) Ghecom, c) Ligsite$_{csc}$, d) Fpocket, e) Depth, and f) AutoSite where the error bars are constructed from the family minima and maxima. Line: $y = x$

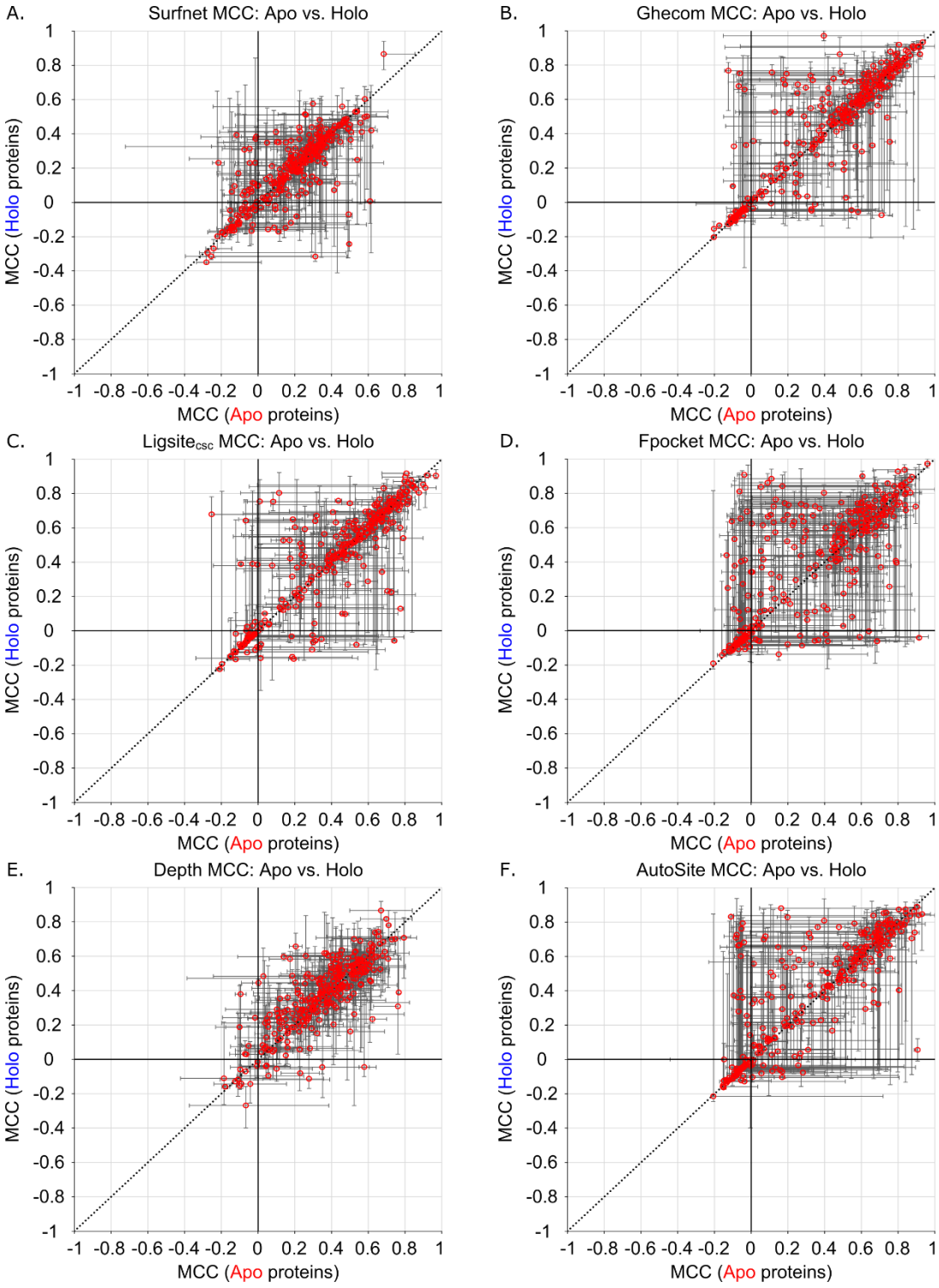


Figure 4-5. Family median MCCs of apo and holo protein structures. Data presented for a) Surfnet, b) Ghecom, c) Ligsite$_{csc}$, d) Fpocket, e) Depth, and f) AutoSite where the error bars are constructed from the family minima and maxima. Line: $y = x$

4.4.3 **Relationship to Structure Quality**

Was it possible that the performance of a given crystal structure in any LBS-prediction method was related to the overall quality of that structure? Structure quality was assessed in two ways: structure resolution and Cruickshank Diffraction Precision Index (DPI).[165] There is some redundancy here, as resolution is used in the calculation of DPI, but DPI is a far more complete measure of X-ray crystal structure quality. Across all comparisons of F score vs. resolution, F score vs. DPI, MCC vs. resolution, and MCC vs. DPI, the highest R^2 value obtained was 0.03. This implies no correlation between structure quality and the performance of the structures in LBS prediction with the methods showcased here.

As an additional metric of quality, unresolved residues were considered. For these experiments, the unified binding sites were examined across all structures within a family, and any missing (unresolved) residues were noted. Residues outside of the binding sites were not tallied in this process. One family was removed from the dataset for due to an excessive number of missing residues in structure 1HNK (see Methods). There were 61 families in the dataset which had at least one structure with at least one UBS residue missing.

The performance of the structures in those families were compared on a per-family basis, i.e. the structures without any missing residues vs. the structures with at least one residue missing. Structure type (apo or holo) was not considered for this analysis. With 61 families, six methods, and 2 performance prediction classifiers per method (F score and MCC), this resulted in 732 comparisons. Of these 732 comparisons, 683 displayed no significant difference ($p > 0.05$) in performance when analyzed using Wilcoxon rank-sum. The remaining 49 cases where there were statistically significant p values represented 19 families.

If unresolved residues were problematic in this analysis, their impact would likely show up in the performance metrics of every method we tested. Instead, 12 of the 19 families only showed statistically significant performance differences for one method, and they were not always the same method: Autosite (3 families), Surfnet (1 family), Depth (3 families), $\text{Ligsite}_{\text{csc}}$ (1 family), Ghecom (4 families), and Fpocket did not show any differences for performance in any family. Of the remaining 7 family cases, 6 of the families showed differences with only two methods, and the last case showed significant differences with five of the six methods.

Most interestingly, the impact on performance of these structures with missing residues is not always negative. The family of Concanavalin A, which showed significant performance

differences for five methods, has three apo structures (1apn, 1dq2, 1enq) which are missing some residues in the binding site and seven apo structures without any missing residues, as well as 10 holo structures which are not missing any residues. The performance of the five methods (Surfnet, Ghecom, Ligsite, Depth, and Autosite) improves on the structures which have missing residues, in every case.

Structures missing UBS residues were uncommon, and those structures causing any significant difference on the performance of any of the methods was exceedingly less common still. As such, we elected to not exclude any of these data, as missing residues appear to have an overall miniscule impact on the performance of the methods.

4.4.4 **Conclusion**

The six LBS-prediction methods showcased in this work mostly failed to accurately predict the full extent of the UBSs in the provided proteins. UBSs are our more robust definition of a protein binding site derived from the bound ligands of multiple crystal structures of the same protein, rather than each structure having its own binding site definition based on a single ligand.

Our UBS definition may be deemed too generous, but they should only aid methods that tend to 'over-predict' binding sites, which is a somewhat expected problem with LBS-prediction methods. This is mostly due to the nature of binary classification methods, as they tend to extract many false positives when being pushed towards a 100% recall rate. The top 10 LBS-prediction methods in round IX of CASP[166] had an average MCC of 0.62 for the 129 targets in that round. While we cannot directly compare to this value, as the datasets and methods being tested are not similar, it does give a reference value for what state of the art methods are capable of in competition. Due to this, the lack of F scores and MCCs with values close to 1 is unsurprising.

More importantly, the predictive power of the six algorithms did not appear to correlate with the ligand-bound state (apo vs. holo) of the protein structure being used. This implies that, contrary to historical belief, apo structures can perform as well as, or better than, holo structures and vice versa. In order to extend this idea to other computational methodology, more high-quality datasets need to be made available to the community which have proper representation of apo structures.

Chapter 5. Protein-Protein Interface Topography

5.1 Abstract

Protein-protein interactions (PPIs) serve as one of the less understood frontiers of drug discovery. These elusive targets present challenges both in their fragility and their physicochemical complexity. Herein, we explore the topographic nature of these supposedly 'flat and featureless' interfaces in attempt to better understand their mechanism of binding small-molecule PPI inhibitors. The protein-protein interaction inhibitors database (2P2I db) serves as a repository for structural data relevant to PPI inhibition, containing data for both the complexed PPI as well as ligand-bound PPI subunits. Data from 2P2I is employed to represent druggable PPIs in a small-scale investigation of the physical and chemical characteristics of PPI interfaces. These druggable PPI interfaces are complemented by a much larger set (derived from PDBbind) of complexed PPIs for which there are no known inhibitors, by which we infer that they are currently less druggable. These less druggable structures are accompanied binding data describing the interaction of their subunits. First, we assess the physical and chemical characteristics of the 2P2I set to determine what makes them amenable to modulation by small-molecules. Second, we apply the same characterization methods to the PDBbind set, to probe for relationships between their affinity of complexation and characteristics that make them more or less druggable. We find that the hollowed areas in druggable PPI subunits are more sharply shaped than other subunits, and they contain primarily nonpolar aliphatic and aromatic residues. Conversely, the protrusions from these druggable subunits primarily contribute charged polar residues to interact with their complexing partner. Lastly, we found no relationship between the binding affinity of the PPI complexes and any of the physical or chemical properties investigated.

5.2 Introduction

Protein-protein interactions (PPIs) represent one of the most complex levels of organization in biological molecules and are the basis of more complicated multimeric proteins.[43] Designing

molecules to modulate PPIs has therefore become a valuable pursuit for medicinal chemists due to their enormous therapeutic potential.[45, 167-168] It became immediately apparent that finding small molecules that adequately modulated PPIs was exceedingly difficult, as the properties of protein-protein binding events are different than protein-ligand binding events.[168] Traditional high-throughput and fragment-based screening approaches have had some success with PPIs, but have only achieved ligands with similar affinity to the natural protein partners.[41] While this appears promising, affinity on the order of the natural complexing partner requires specific structural circumstances in order to outcompete the energetically-favorable, hydrophobically-collapsed state of the complexed PPI.[41] PPIs may be characterized by a number of descriptors including: obligatory nature, binding strength, stability, number of interacting partners, and shape.[43, 51] Here, we revisit the basis of these descriptors, as understanding physicochemical principles of PPIs is a crucial first step in determining their druggability as a target class.[167]

5.2.1 Characteristics of PPIs

5.2.1.1 Obligation

The obligation of two protein partners to complex with each other is defined by the necessity of their complexation in order to carry out their intended function *in vivo*. Obligation is derived from a biological context, where some proteins are complexed immediately after synthesis or during protein folding.[167] Obligate PPIs are therefore thought to be permanent interactions. Permanent PPIs are so stable that attempting to dissociate the protein partners will result in denaturing one or both partners.[167] This property makes these permanent PPIs an extremely difficult set of targets for drug design, as any potential drug molecules would need to act on these targets upon their synthesis or immediately thereafter during folding and complexation.[167] Even though transient PPIs are therefore theoretically more amenable to modulation by small molecules, some still consider them undruggable by traditional medicinal chemistry approaches.[45]

The obligate nature and duration of complexation between two protein partners are not binary classifiers. Instead, there is a continuum between non-obligate and obligate, as well as transient and permanent interactions when describing protein-protein partners.[40] The nature of obligation between two protein partners would intuitively be coupled with the binding affinity which describes their interaction, but this correlation has been disproven.[40] That said, binding affinity has been used to define and separate transient complexes from permanent complexes.[169]

5.2.1.2 Stability

Immediately related to the permanence of a protein complex is the stability of the complex. Protein-protein complexes are typically either extremely stable or conversely unstable. In the case of stable complexes, the stability of the complexed set of protein partners is a mirror of the instability of the individual partners upon dissociation. That is to say, in these extremely stable protein-protein complexes, at least one of the partners is unstable when the complex is broken.[167] In the case of unstable complexes, each of the protein partners is stable by itself in an uncomplexed state, implying less favorable pressure for complex formation. It has been suggested that proteins which participate in these weakly transient associations are more likely to be promiscuous binders, participating in numerous PPIs.[168, 170] Crucial work by Kastritis *et al.*[171] probed stability of PPIs through testing the sensitivity of PPI affinity to the following environmental factors: pH, ionic strength, temperature, presence of small molecules, and covalent modifications (such as phosphorylation). Their findings suggest that the stability and affinity of a protein-protein complex can be moderately affected by temperature and ionic strength and exceptionally affected by pH.[171] These findings shed light on the ever-difficult task of acquiring structural information for PPIs, as these environmental factors are the very tool that crystallographers utilize to grow protein crystals. With this in mind, it is unsurprising that either the complexed or uncomplexed states of many PPIs remain structurally elusive.

5.2.1.3 Multimeric States

PPIs can describe many different multimeric protein states, from simple binary homodimer complexes all the way to unpredictably chained oligomeric interactions, such as those found as amyloid β-peptide fibrils, indicated in advanced stages of Alzheimer's disease[172]. Multimeric interfaces which incorporate more than two partners complicate every aspect of drug design. For this reason, our work along with many past studies concentrates on binary complexes.

5.2.1.4 Shape

The shape of a PPI can be described on multiple levels. On the tertiary level, the entire PPI has a given shape which falls into one of a few categories: flat, engulfed, twisted, or armed.[43] Flat interfaces are the simplest type, with no distinct structural characteristics protruding across the interface (Figure 5-1A). Flatness of a PPI surfaces is characterized by its planarity, a mathematically determined parameter.[40, 43, 47] Engulfed interfaces describe an interaction where a

small globular protein is complexed with another partner (which is typically much larger in size), which engulfs it (Figure 5-1B). Twisted interfaces contain structural elements that twist together, commonly beta sheets or alpha helices separated by loops from the bulk of the globular protein (Figure 5-1C). Armed interfaces exist where there are protein extremities such as loops or even small domains which wrap themselves around their complexed partner(s) (Figure 5-1D). Jones & Thornton[47] suggested that heterodimeric complexes had a tendency to be flatter, and more planar, while homodimeric (more permanent) complexes often displayed more twisted, engulfed contact surfaces. Other works have noted that PPIs are relatively flat and featureless as a whole[41, 44-46], suggesting that designing small molecules that bind efficiently to PPIs is extremely difficult.

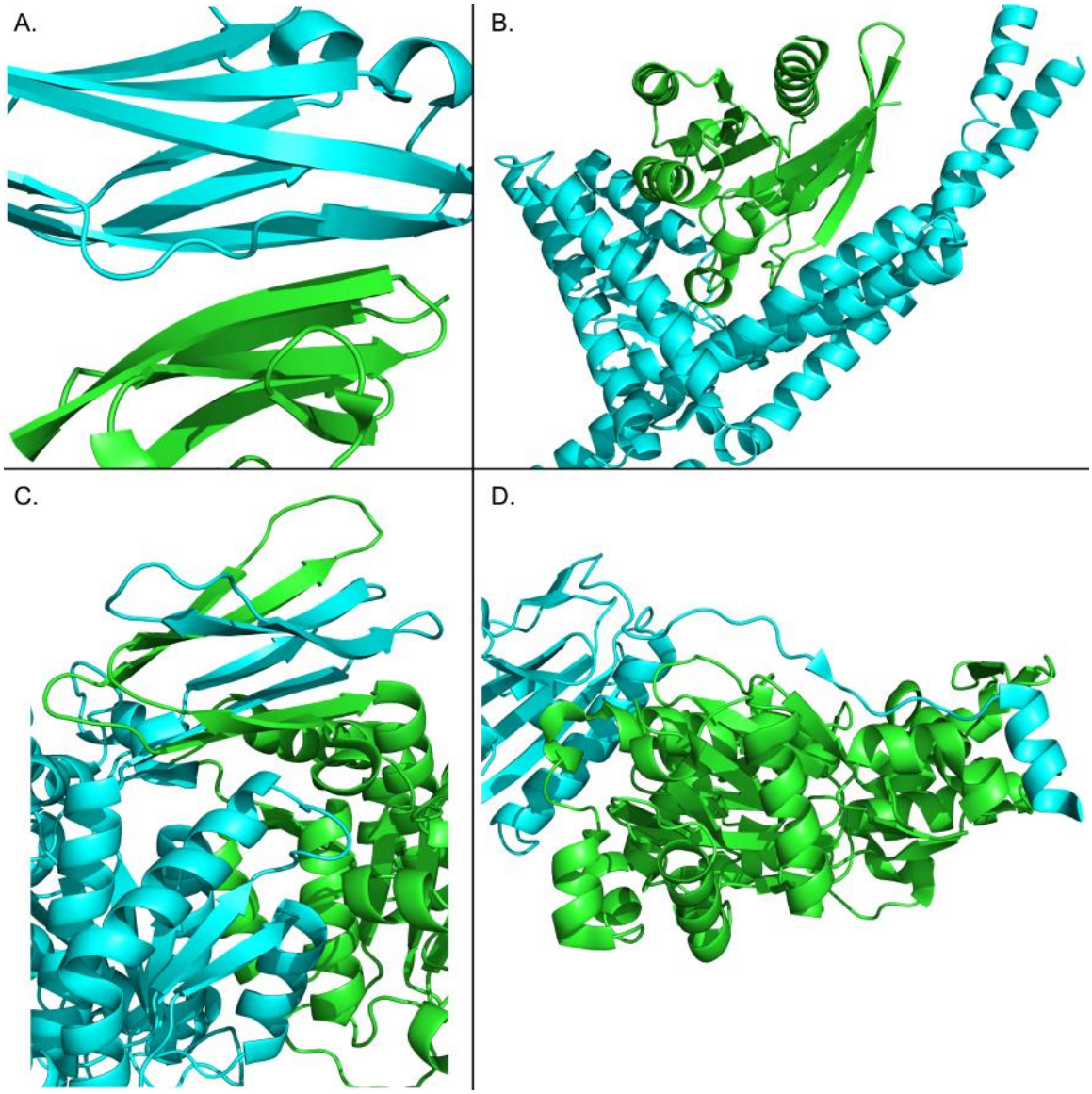


Figure 5-1. Different types of interfaces. The four general geometric varieties of PPIs: A) Flat (PDBid: 2WP3), B) Engulfed (PDBid: 3TNF), C) Twisted (PDBid: 5FYN), and D) Armed (PDBid: 1AHE).

5.2.2 Differences from Protein-Ligand Binding

First and foremost, PPIs describe events whereby a protein interacts with another protein chain instead of a small-molecule partner, so there are no natural substrates we may start from for structure-based drug design.[41] As a reflection of that protein character, many early stage inhibitors tend to be peptide-based, which have drawbacks for pharmaceutical purposes.[168] Peptide inhibitors tend to have a higher molecular weight than traditional small molecules, drastically decreasing ligand efficiency and also resulting in poor bioavailability.[168, 170]

Ligand recognition, for terms of specificity, is governed largely by shape complementarity of a ligand to the protein target.[168] PPIs archetypally lack any deep pockets[44] and are thought to be flat and featureless[41, 44-46]. These characteristics present enormous specificity problems for potential inhibitors, not the least of which is finding inhibitors that modulate PPIs in the first place. There is an underrepresentation of known PPI inhibitors in the datasets used to train *in silico* methodology used for screening purposes.[44] Due to this, discriminating molecules capable of modulating PPIs from drug decoys is difficult.[44]

Energetically, protein-ligand interactions (PLIs) and PPIs are both largely driven by the hydrophobic effect.[173] However, protein-ligand binding sites are typically small enough that the ligand by itself can aid in displacing the bound waters within the site to drive the hydrophobic effect. The majority of protein-protein interfaces are primarily hydrophobic in composition, with patches or zones of hydrophilic groups at the outer edges of the interface in order to help draw water out of the interface upon complexation.[43, 53] Therefore, in order for a ligand to bind to a PPI interface with the same mechanism of action as a typical orthosteric inhibitor, the ligand would need to be nearly the size of the interface. If the energetically beneficial properties of protein-protein complexation were distributed evenly about the entire interface, modulation to be dissociated by a small molecule ligand would likely be impossible. This cannot be the case, as small molecule inhibitors of PPIs do exist, begging the question "What is their mechanism of action?".

Successful binding of a small-molecule ligand to a PPI implies that there must exist a location specific to that binding event, due to the size difference between the effector and the target. Indeed, the presence of a defined pocket capable of binding small-molecule ligands is

necessary for drug-class modulation of a PPI, but that alone is not sufficient.[174] Moreover, PPIs utilize a larger number of smaller pockets compared to PLIs, where fewer, large binding pockets normally observed.[45] As small molecule ligands have been demonstrated to destabilize and dissociate PPIs, the energetically favorable characteristics of PPIs are therefore not evenly distributed.

PPI interfaces are large, but there are crucial, smaller regions —recognition patches, or hotspots— which are responsible for the majority of the free energy of binding.[167-168, 175-176] The average interface area of homodimers has been measured at ~1940 Å^2, and heterodimers at ~970 Å^2. These total interface areas can be clustered into smaller contact patch regions which may function as recognition patches.[54] The first (largest) patch in any interface is typically on the order of 800 Å^2, with every subsequent patch being much smaller.[54] There is a correlation between interface size and the number of hotspots present in the interface, where larger interfaces have more hotspots.[53-54] Additionally, homodimers also tend to have more, larger hotspots than heterodimeric complexes.[54] Identifying hotspot residues through computational means is relatively difficult, but sequence mutagenesis studies can reveal crucial hotspot residues through noticeable drop in the free energy of binding.[171]

5.2.3 **Definition of PPI Interfaces**

Multiple protocols have been used for defining which atoms or residues are considered part of a PPI interface. Some of the earliest work defined the interface residues as any displaying lower solvent-accessible surface area (SASA) in the complexed dimer when compared to the lone monomers.[51, 177] This method was augmented with thresholds for residue-level and atom-level definitions[43] as follows: residues considered part of the interface decrease by > 1.0 Å^2 upon complexation, and interface atoms decrease by >0.01 Å^2 upon complexation. These thresholds were incorporated to the method to account for crystallographic errors for atom placement and small computational errors when calculating SASA.[40, 43, 47] Defining the interface atoms *via* decreases in SASA upon complexation should yield the locations where water has vacated as the interface halves collapse within each other, thus representing the role of the hydrophobic effect.

Predicting PPI surfaces and residues became a logical next step in the characterization of these critical biochemical events. During the 1990s-2000s, protein-ligand docking exercises as well as ligand-binding site prediction methodology were readily benchmarked and quickly became

community tested and better understood as tools for drug design and other applications. [178] Until that point, protein-protein docking software had been pursued but only really existed in closed academic environments. As docking and prediction software go hand-in-hand, benchmarking and community-encouraged development was necessary for growth of these technologies.

The Critical Assessment of Predicted Interactions (CAPRI) is a community initiative dedicated to the prediction of protein-peptide and protein-protein interactions from unpublished structural data, introduced in 2002.[178] Inspired by the community-wide impact that CASP[179] had on pushing the limits of protein-folding technology, CAPRI is crucial to driving creators of prediction functions, as well as scoring functions to adapt and improve their methodologies to cope with the robust-nature of PPIs. In the most recent, 6th round of CAPRI, the authors note tangible improvements in the best state-of-the-art methodology for predicting protein-peptide complexes, and low-to-medium quality protein-protein complexes.[180] Still, the authors conclude that adequate modeling of conformational flexibility between interacting proteins needs improvement.[180] The authors also note that publicly available webservers for PPI prediction underperform private algorithms in the task of predicting protein-peptide interactions, owing to this being a newer class of targets and the webservers being undertrained on them.[180]

5.2.4 **Datasets and Databases**

A common problem with past research in computational biology is the transparency and availability of datasets used. Early publications did not have the luxury of pre-curated databases for specific purposes such as PPIs or PLIs and thus, had to curate their own datasets. The primary issue with this is comparison of results across multiple studies. While studying new targets is imperative for discovery of new information, revisiting previously studied targets serves as an invaluable means of checking newer methodology where reference material exists. A number of useful datasets and databases have been curated for the purposes of small molecule discovery in the context of PPIs. Some are dedicated specifically to this cause, and some are subsets of larger databases.

5.2.4.1 ***2P2I db***

The protein-protein interaction inhibitors database (2P2I db[55-57]) is a collection of structural data for PPIs where there exists at least one high-quality structure for the complexed protein

partners and at least one high-quality structure of one of the protein partners bound to a small molecule or inhibitory peptide. This collection contains only hetero complexes and does not contain covalently-bound inhibitors, as both of those cases exhibit wildly different behavior. The inhibitors in this set are also of orthosteric nature, as allosteric mechanisms would occur outside of the protein interface.

PPI structures are cultivated from the PDB[58] using the Dockground[181] server. Filtering is accomplished by keeping only structures with resolution > 2.0 Å^2, discarding disordered proteins and complexes with nucleic acids. This yielded 202 heteromeric complexes for the initial version of 2P2I db. The PDB is then queried using an advanced query to obtain free protein structures corresponding to each complex bound to small molecule inhibitors. Inhibitors are manually checked to ensure the binding location is indeed at the interface. The end result is nine protein-protein complexes and 25 protein-ligand complexes for the first published version of 2P2I db.[56] The most recent update (accessed: 7/09/2018) contains 31 PPIs and 242 small molecule inhibitors.[57]

These proteins are divided into three classes: protein-peptide complexes (class 1), globular proteins (class 2), and bromodomains (class BRD). Many of the structures of the bromodomains are constructed from homology models (for BRD3-1, BRD3-2, BRD4-2, BRDT-1, and KRAS) using closely related homologues as templates (identity ranging from 75-94%).[57] Binding data are acquired, where available, from Binding MOAD[3], PDBbind[61], or BindingDB[72]. The update process for this database is largely automated, so updates are frequent.

The nature of 2P2I's curation emphasizes integrity of structural data with respect to known binders, as information about ligands is only provided where there exists structural data to accompany it. While having additional ligand data seems appealing, the biochemical accuracy of the data, i.e. matching chemical conditions, is imperative to the relevance of the data. Therefore, excluding extraneous data which is not coupled directly to structural information is an appreciated characteristic of this data resource.

5.2.4.2 ***PDBbind***

The curation and contents of PDBbind are extensively discussed in section 2.2 of Chapter 2.[61] Along with the wealth of protein-ligand data that PDBbind possesses, it also contains a downloadable set of PPI complexes. Notably, there are also binding data for many of these complexes. For instance, the 2016 release of the protein-protein complex dataset had 1335 K_d data,

129 K_i data, and 37 IC_{50} data. Interestingly, the more useful and accurate K_D measurements are far more common for these difficult PPI targets. Despite advertisement as primarily a protein-ligand database, this is one of the largest collections of curated protein-protein affinity binding data.

5.2.4.3 TIMBAL

TIMBAL[182-183] is a ligand-centric database, focused on holding molecules with molecular weight <1200 Daltons that modulate PPIs. Peptide molecules are limited to 10 peptide bonds, or 11 total residues. The database covers 50 known PPI drug targets, including protein complexes that can be stabilized by small molecules for therapeutic effect. Originally hand curated, the newest updates of TIMBAL are exclusively automated searches of ChEMBL[184]. The ChEMBL data are then matched with structures from the PDB to obtain protein-small molecule, protein-protein complex, and unbound protein structural data. The newest version (update 10-JUN-2015 using ChEMBL_20) of TIMBAL contains >14,000 data points for ~7000 small molecules with 50 PPI targets.[183] Notably, more than 9000 data entries are for integrins alone, as the surface receptors have been pursued as therapeutic targets for nearly two decades. The small molecules are separated into inhibitory and stabilizing classes, the latter being a small subset of the data. However, there are therapeutic opportunities for molecules with such mechanisms of action.

The size of this collection is impressive but harvesting binding data purely from ChEMBL and matching them to structural data later is not appropriate if care is not taken to match environmental conditions such as pH and temperature. Use of this database should be accompanied by rigorous data validation to ensure that binding conditions and structural conditions are appropriately matched. Importantly, there is no requirement that the inhibitors included in TIMBAL's dataset be orthosteric modulators, so some allosteric inhibitors are likely present. This may be a boon or a hindrance, depending on the experimental context.

5.2.4.4 iPPI-DB

The iPPI-DB[185-186] is another ligand-centric database. The development of iPPI-DB was focused on curating a collection of low molecular weight compounds capable of modulating the function of PPIs. The authors describe low molecular weight as compounds with MW < 1000 g/mol. Entries are curated directly from journal articles and patents. The authors claim only journals with "expertise in medicinal chemistry" are considered but provide no definition of how this condition is applied to their curation process. They require that the PPI targets be discussed in

multiple publications in order to stress targets that are well documented to have implications in a disease state. Valid ligands are only allowed to contain the following atom types: C, N, O, S, P, and halogens. Peptides and macrocycles are discarded, though the authors provide no details as to the description of these molecules for their filtering process. The latest publication of iPPI-DB contains 2461 binding data for 1650 compounds across 13 families of highly homologous PPI targets.[186] The most recent version of their dataset (accessed 7/05/2018) contained 1756 ligands across 18 families of PPIs.

While iPPI-DB's web interface is impressive and user friendly, there are major concerns about the vague descriptions of the cultivation of their dataset. The database is forwardly ligand-centric, with queries focusing specifically on the types and affinities of ligands the user aims to locate, also allowing for designation of general PPI homologous family of interest. Structural information is not provided for their protein families, and individual homologues are not available as searching options. This relatively relaxed classification of their protein targets is alarming, given the strong language used to confer their commitment to traditional medicinal chemistry approaches and relevancy of molecules and the targets they bind to. The authors also do not stipulate the mode of PPI modulation, so allosteric ligands are likely also part of their dataset.

5.2.5 Relevant Webservers and Tools

Many computationalists prefer to make their own in-house tools, but some more complicated methodology has been incorporated into webservers and publicly available tools. The most significant contributors have been noted below.

5.2.5.1 2P2I inspector

Along with their comprehensive database discussed in 5.2.4.1, 2P2I also contains a useful tool for computing various physical, chemical, and geometric properties of PPIs. Interface propensities are calculated using in-house shell and tcl scripts[57], and presented using VMD (http://www.ks.uiuc.edu/Research/vmd). Missing hydrogen atoms are added with PyMOL (https://www.pymol.org). Gap volumes, planarity, eccentricity, and circularity are computed using the PRINCIP program imbedded in SURFNET[124].

5.2.5.2 PROTORP

PROTORP[187] was a web server dedicated to the analysis of protein-protein interactions in 3D structures. The server calculated a number of physical and chemical parameters related to the binding energy of association for the two components of the complex. Some parameters include: residues present in the interaction site, size of the interaction site as measured by SASA (calculated with NACCESS), and geometric characteristics including planarity, length, breadth, and numerical eccentricity. Gap volumes are also provided. While ProtorP is no longer online, it was significant web resource in its time. Many aspects of the ProtorP server were calculated using the PRINCIP program within SURFNET[124], as it was developed by the same group.

5.2.5.3 InterProSurf

InterProSurf[188] is a web server dedicated to predicting which surface residues on a provided protein are most likely to interact with other protein targets. The prediction method is based on SASA of the interface residues combined with known propensity values (from past publications) for PPI interfaces and a clustering algorithm to locate patches of surface residues which match the appropriate characteristics of a protein-protein interface.

5.2.5.4 PDBsum

PDBsum[189-190] is a web server which provides structural information on entries of the PDB. The analytics performed by PDBsum are provided in a very visual format, ranging from structural characterization of secondary structure to analysis of structural quality and details of protein-protein contact regions. PDBsum is a notable resource for target-specific study of various PDB structures. The image-based output of the server is not amenable to cultivation of data for large datasets, but the represented information is available in text format through web utilities such as *wget*. The entire contents of the PDBsum database are downloadable, as well. While this is not a dedicated resource for PPI information, it contains a dedicated tab of information for PPI structures. The PPI interface residues can be extracted from the protein-protein contact plots.

5.2.5.5 PDBePISA

The PISA server (Proteins, Interfaces, Structures and Assemblies), also termed PDBePISA, is a web tool provided by the European Bioinformatics Institute (EBI) as an extension of the European Molecular Biology Laboratory (EMBL). The PISA server analyzes interfaces found in

asymmetric units and predicts probable quaternary structures based on crystallographic data. The tool is tied to a database, which contains pre-calculated information as to make analysis much quicker for users. Many terms and parameters are described as provided on the website, but the methods used for calculation are not provided, and there is no overarching publication in which to find them. Some provided parameters for an input PPI structure include: surface area of contact patches within the provided structure, $\Delta G_{solvation}$ as the solvation free energy gain upon formation of the specific interface being analyzed (many are provided and labelled accordingly), symmetry operators of the interface, number of interacting atoms in the interface, number of hydrogen bonds, number of salt bridges, number of disulfide bridges, and some very customized parameters related to the probability of the formation of that interface, as well as its significance.

While the information provided by the server is undoubtedly useful for the study of individual proteins of interest, and the data are downloadable, users cannot process many structures at a time. This renders the tool less useful for large-scale analyses of datasets containing many protein structures.

5.2.6 **Investigations of Physicochemical Properties of PPIs**

A plethora of previous publications have studied physicochemical parameters of PPIs including: residue composition, SASA, accessibility of side chains, electrostatics, and localized geometry of the interface residues. As the studies occurred over a large span of time, they will be addressed chronologically.

5.2.6.1 Early Work

Early work by Janin *et al.* in 1988[51] used a set of 23 oligomeric proteins (16 PDB structures, 7 donated from cited laboratories) to investigate amino-acid composition, hydrophobicity, and SASA properties of the PPIs. The most important contribution of this work was the introduction of the SASA-based method of determining PPI interface residues. Residues displaying any degree of -ΔSASA when going from their uncomplexed to complexed state are considered to be part of the interface. Other significant findings include: smaller interfaces have approximately 700 $Å^2$ of buried SASA, while larger ones cover between 3000-10000 $Å^2$, which constitutes up to 40% of the protein's total surface area in the extreme case of catalase. These buried SASA figures were calculated with the following equation:

$$SASA_{interface} = \frac{(SASA_{subunit\ 1} + SASA_{subunit2}) - SASA_{complex}}{2}$$

where $SASA_{subunit1}$ and $SASA_{subunit2}$ are the results of calculating the SASA of the isolated subunits, with the assumption that their structure does not change upon separation. Oligomers with small interfaces had globular subunits with accessible surface areas similar to those of monomeric proteins, which suggests that those small-interfaced dimers assembled from pre-formed protein monomers. Some data about the composition of the interface residues vs the general solvent-exposed protein exterior, as well as the interior (core), were gathered. However, this dataset was very small, so development in this area required more data for clarity.

Similar work by Argos was completed in 1988[177] with a dataset composed of 58 proteins with 24 oligomeric subunits and 34 independent, uncomplexed domains. This study concluded that the composition of protein-protein interfaces was similar to that of the general protein surface. Interestingly, they found that the contact surfaces showed a relatively uniform distribution of polar and non-polar atoms over the interface surface. Later works find the opposite of this, so this finding was likely due to the small dataset. There was an increased abundance of large, aromatic amino acids, as well as Arg present in the interface when compared to the general protein surface. There were also an unusually large number of self-contacts (i.e. Met-Met, Phe-Phe, Leu-Leu, and Asn-Asn), as well as less-traditional pairings (Ser-Asp, and Tyr-Asp) within in the interfaces. They found that, on average, subunits and domains lose approximately 20% of their water-accessible surface area to form the respective interfaces, with considerable variation (5-43%). This was one of the first studies to investigate the geometric orientation of the interfaces, using complicated spline-interpolations on a 0.5 x 0.5 Å grid. The interface surfaces often displayed a flat overall cross-section, especially for subunits where symmetry is involved. Domain interfaces were nearly evenly divided between flat and curved. Finally, Argos found that ~70% of the total interface surface was contributed by single residues in distinct structural units, which is to say that the interface is not represented by a continuous chain of amino acids, but rather fragments of distant motifs.

5.2.6.2 Thornton's Major Contributions

Jones & Thornton published a robust article[43] on similar topics and introduced interface planarity as a concept in 1995. This study used a dataset of 32 nonhomologous protein dimers

cultivated from the PDB. This was an exhaustive study of physical, chemical, geometric, and propensity properties of PPIs. Four metrics for assessing PPIs were discussed therein.

First, to better quantify the geometric shape and orientation of the interface, planarity was introduced. Planarity was defined by the RMS deviation of all atoms in an interface from the least squares plane through those atoms. Furthermore, the circularity of the plane was also considered, as a ratio of the standard deviations along the *x* and *y* dimensions of the defined plane (within the plane). The circularity of the proteins in this study ranged from 0.48-1.0 for 31 of the 32 proteins, with the last yielding a circularity of 0.33. They noted that the oblong shape of this final protein was likely due to its biological function as a parasitic surface coat protein.

Second, the concept of interface segmentation was introduced in order to better understand Argos' finding where the interfaces consisted of fragments of distant structural motifs. Interface residues separated by more than five residues in the sequence were allocated to different segments to help quantify the degree of segmentation in the interfaces. The number of segments in their datasets ranged from 2 to 15 and there was a weak correlation with the size of the interface (r = 0.59).

As a third addition to classifiers, gap volume —a parameter implemented in SURFNET[124]— was further conceptualized into a gap volume index to quantify the complementarity between the surfaces of the two subunits which compose a PPI. Gap volume is calculated by inserting spheres with up to a 5 Å radius between every pair of subunit$_1$-subunit$_2$ interface atoms and shrinking the volume of the sphere gradually until no other interface atoms are contained within the sphere. If the radius is shrunk below a given threshold (default 1.0 Å), it is removed completely. Remaining gap spheres are summed across the entire interface to yield the total gap volume. This gap volume index is then calculated as: $Gap\,Volume\,Index = \frac{Gap\,Volume\ (\text{Å}^3)}{Interface\,SASA\ (\text{Å}^2)}$. The gap index values for their 32 proteins revealed that homodimers and permanent heterodimers tend have higher surface complementarity than transient heterodimers. This should render transient complexes as more druggable, but it is important to note that this surface complementarity is merely a measure of packing density in the interface and does not reveal any information about the interaction network holding the interface together.[167] Parallels were drawn to strength of binding seen *in vivo* for homodimers where some protein complexes will denature before they dissociate.

Lastly, protrusion of interface residues from the molecular surface of the interface were measured using SASA values for the twenty standard amino acids, comparing the interface residues to the generic protein exterior residues. Their findings showed that all residues, except for Ser, showed increased accessibility in the interface (mean exterior residue: 36.09 $Å^2$ vs 43.56 $Å^2$ for interface residues), some showing large increases (Trp, Tyr, Phe, Arg, Met) of ≥20% more exposed surface area. The decrease in accessibility for Ser in interfaces was minimal (~1% relative surface area decrease). From these results, they concluded that the increase in SASA inside the interface points to increased flexibility and side chain mobility inside of the interface. The implication of such a flexible state could indicate that interfaces are not preformed but that residues take on new conformational states when/after dimerization takes place.

Importantly, the SASA-based determination of PPI interface residues first implemented by Janin[51] was utilized in this work, but altered slightly. The authors added a threshold whereby, to be considered as part of the interface: single atoms must display ΔSASA < -0.01 $Å^2$ and residues must display ΔSASA < -1.00 $Å^2$ when going from the uncomplexed to complexed protein state. Additionally, the ending notes stated that many of the features calculated in the publication were to be made available as a computational tool in the near future. The PRINCIP program within SURFNET[124] suite is that later product.

Additional findings include: SASA is roughly correlated to molecular weight of protomers (r = 0.69), buried SASA per subunit ranges from 368.1 - 4746.1 $Å^2$, there are an average of 0.88 hydrogen bonds per 100$Å^2$ of SASA buried (for interfaces covering > 1500 $Å^2$ per subunit), and interfaces are more hydrophobic than the exterior of proteins but not as hydrophobic as protein cores.

5.2.6.3 Janin's Major Contributions

Three studies in 1999-2003[52-54] were completed by Janin and coworkers with datasets of 75 heteromeric protein-protein complexes for the former two and an additional dataset of 122 homodimers for the final study. With the similarity of the three studies, their findings will be discussed jointly. The geometric model of an interface composed of rim residues and buried residues is heavily discussed, where any interface residues (defined by those that lose SASA upon complexation) that still have SASA after complexation are rim residues, and those that approach ~0 $Å^2$ of SASA are buried residues. The authors discuss some metrics using "buried SASA"

referring to the SASA which is no longer present on complexation of the two subunits, therefore burying it.

The authors clustered the atoms in their interfaces using an average linkage clustering method in order to subdivide the contact surfaces into patches, also referred to as hotspots. The average homodimeric interface contains one or two patches. Each of these patches bury a total of 600-1600 $Å^2$, are composed of 65% nonpolar atoms, and include ~18 hydrogen bonds. About 77% of the homodimers in their dataset had a buried SASA > 1000 Å. The typical heterodimeric complexes had a buried SASA between 600-1000Å. They found that homodimeric interfaces were typically twice as large as heterodimers, on average. Each side of a heterodimeric interface has an average of one hotspot, which is about 800 $Å^2$ in size. Homodimers tend to have more hotspots which are larger. The cores of the interfaces, composed of the previously defined "buried" residues, were an average of 32 residues per monomer in homodimers, and only 12 residues per component of their heterodimeric complexes. The core residues of the interfaces were enriched in aliphatic and aromatic residues, depleted in charged residues (with the exception of Arg), and generally resembled the interior of the protein. Despite these contrasting figures between homodimers and hetero complexes, they both contained approximately 1 hydrogen bond per 75 $Å^2$ of polar interface area.

5.2.6.4 Binding Affinity of PPIs

The most recent publications on PPIs have not discussed as many of the geometric characteristics of their binding state and have concentrated more on other issues such as stability and relationships to ligand binding events.

A study by Kastritis *et al.*[171] in 2011 investigated the effects of environmental factors such as pH and temperature on the stability of PPIs as measured by changes in their $\Delta G_{binding}$ with a dataset of 144 protein-protein complexes curated from the PDB. This dataset included antigen-antibody pairs, enzyme-inhibitor, enzyme-substrate, and enzyme-receptor complexes, as well as G-protein pairs. All structures were coupled with K_d measurements ranging between $10^{-5} – 10^{-14}$ M. This study exercised rigorous attention to detail in terms of the binding affinities used. While the binding affinities were not necessarily included alongside the structural information in the publication of origin for each structure, the authors of this work painstakingly manually curated binding constants appropriate for the crystallization conditions for each target structure in their dataset. They also considered the biophysical techniques utilized to acquire the binding constants

and the associated error metrics that could result from the measurements, as well as binding constants achieved as a conversion (i.e. calculation of K_d from an appropriate K_i).

The authors found that the measurements in their dataset had all been conducted from 18-35°C, with only 3 exceptions. The potential impact on the affinity of the PPI complex from this was small. They found that changing ionic strength in the range of 0.1—0.5*M* could have a similar, but slightly larger effect. However, pH was undoubtably the most powerful environmental variable for controlling the affinity of a PPI complex. Change of pH in the range of 5.5-8.5, which covered about 95% of their dataset, resulted in a change of K_d by a factor of 10-50, corresponding to $\Delta G \sim$ 1.4-2.3 kcal/mol. The impact of pH surpassed the effects of temperature and ionic strength by far.

They noted that collecting affinity data from publications was exceedingly laborious and had confidence in their paired affinity-structure pairings to within a factor of 2-10 for K_d, or 0.4—1.4 kcal/mol for ΔG. Most importantly, the concluding remarks were aimed towards prediction and modeling of binding affinities for PPIs: "… it makes little sense to model or predict a K_d to within better than an order of magnitude, unless one is also prepared to model its dependence on pH, and possibly also ionic strength and temperature."[171] Notably, this becomes even more crucial for the elusive low-affinity complexes, for which obtaining structural data is exceedingly difficult due to their fragility when preparing crystals.

5.2.6.5 Drug-like Ligands and Small Molecule Properties

A study by Karanicolas and coworkers[44] aimed to characterize physicochemical differences between PLIs and PPIs utilizing the Astex diverse set[129] and a PPI dataset of 21 nonredundant complexes derived from TIMBAL[182] and 2P2I db[55]. To create their PPI set, they filtered complexes contained in TIMBAL and 2P2I db for only those containing non-covalent, orthosteric ligands between 200-675 Da. In cases where more than one inhibitor-bound structure had been solved, only that with the tightest binding affinity was kept. Structures containing cofactors were excluded. The Astex diverse set was used as a starting point for their "druglike" PLI dataset, but structures containing cofactors or secondary ligands were removed. Similarly to the PPI set, complexes with ligands outside of the 200-650 Da range were removed (as the largest in the PPI set was 651 Da). The final PLI set consisted of 46 binary protein-ligand complexes. A screening portion of the experiments also utilized a drug decoy set of 10,000 randomly selected compounds with MW between 200-750 Da from ZINC[191]. The DUD-E server[192] was also utilized to create a custom-tailored set of 50 decoy compounds for their targets.

The overall ligand efficiency of PPI-bound ligands was lower than those of PLI-bound ligands. Bound inhibitors at protein interaction sites retained more exposed surface area than their counterparts at PLI sites. This also held true for other analogous sets of drug-like complexes, such as the full DUD-E set[192]. They noted that, of the traditional drug targets, serine proteases exhibited the most exposed bound-ligand surfaces. Virtual screening experiments utilizing the FRED software package[193] were intended to be relatively easy on the screening software and aimed to probe their performance on PPI targets. For the more traditional Astex-based set, the known inhibitor was ranked within the top 2% of the library in 80% of the targets. Conversely, only 50% of the targets in the PPI set had their known inhibitor ranked within the top 2% of the library. This merely illustrates the bias our current methodology has towards protein-ligand binding events in typical, orthosteric ligand-binding sites in proteins, as opposed to the relatively exotic binding locations of protein-protein interfaces. Lastly, they docked some of the highest ranked decoys from the PPI screening experiment and observed the fraction of exposed ligand surface area (Θ_{lig}) in the bound pose. They found that the highest ranked decoy molecules also displayed high Θ_{lig}, much like the known PPI-inhibitors, despite this characteristic being more unique to PPI inhibition when contrasted against typical targets for small-molecule inhibition. They concluded that this observation pointed to protein conformation being the primary determinant of Θ_{lig} in inhibitory complexes, furthering the evidence that the physical landscape of PPIs is very different from traditional drug targets.

5.2.7 **Moving Forward**

While many aspects of PPIs have been heavily investigated in the past, experts deemed that structural characterization parameters were inadequate to differentiate between different affinities or specificities of diverse PPIs at the time (2002).[40] The recent virtual screening study by Karanicolas and coworkers[44] furthered the idea that the physical landscape of PPIs is different than PLIs. More investigation and quantification of those differences is necessary. Issues of data availability for PPI targets have been lessened with the emergence of well-maintained databases devoted specifically to curating data for PPIs and PPI inhibition, and the new-found presence of structural data for PPIs in both their complexed, and ligand-bound states. For these reasons, revisiting ground-level physical characterization of PPIs seems appropriate again at this time.

We aim to investigate the localized physical and chemical properties of druggable protein interfaces by examining a dataset of druggable PPI interfaces derived from 2P2I db[57], where there

are structural data representing both the complexed PPI state and a ligand-bound, inhibited subunit. After establishing the characteristics of these druggable interfaces, we investigate a dataset of less druggable PPIs derived from PDBbind[61]. The proteins in this less druggable PPI set have no known inhibitors but binding data describing their affinity of complex formation is known. These less druggable structures will be separated into three data subsets based on their binding affinity, and their physical characteristics will be assessed for any patterns matching those of the druggable interfaces.

The physical characterization will involve fitting planes to their interfaces, similar to the approach utilized by Thornton and coworkers.[40, 43, 47], clustering various groups of atoms in relation to those planes, and quantifying structural features of protrusions and hollows that are present within the interfaces. By clustering the contact atoms in the interfaces, we can also extract information about which residues are most crucial, based on their positions relative to the subunit they belong to. Characterizing the local topography of these druggable and less druggable PPIs will ultimately allow us to assess whether or not these interfaces are truly flat and featureless.

5.3 Methods

5.3.1 Dataset Acquisition and Filtering

Two different overarching datasets are utilized in this work, an adaptation of the PDBbind PPI set[61] and an adaptation of the 2P2I dataset[57]. Their adaptations and subsets are noted below.

5.3.1.1 PDBbind PPI set

The PDBbind PPI set serves as the basis for our complexed PPI dataset, intended to contain targets with no known small molecule inhibitors for either subunit in the complex. As ligands are not involved in this dataset, structural information coupled with protein-protein binding affinities was desired as an added layer of analysis.

The 2016 version of the PDBbind PPI set was acquired from their website: (www.pdbbind.org.cn). The 1335 structures coupled to K_d or pK_d data were chosen as a starting point. The Uniprot[81] codes for all protein chains in the dataset were acquired. Any structure containing a chain with a Uniprot sequence found in the 2P2I[57] dataset (accessed 5/2016) or the TIMBAL[183] dataset (accessed 5/2016) was removed, as existing in either dataset implies existence of a known inhibitor. Non-X-ray structures and structures with resolution > 2.5 Å were removed.

Structures with interfaces involving more than two unique protein chains, as well as structures with DNA/RNA complexes were removed. Homodimer complexes were also removed, as the targets in the 2P2I set are all heterodimers. The resulting PDBbind set consists of 347 PDB structures of unique protein-protein pairs with corresponding affinity data measured as a K_d.

5.3.1.2 2P2I Dataset

The 2P2I dataset serves as the basis of our "druggable" PPI set. In this set of druggable of PPIs every complexed pair of protein subunits is complemented with a ligand-bound form of at least one of those subunits. The 2016 release (accessed 5/2016) of the 2P2I[57] dataset was acquired from their website: (http://2p2idb.cnrs-mrs.fr/).

Each PPI complex represented in this dataset was required to have at least 1 complexed PPI structure and one ligand-bound PPI structure after our filtering of the dataset. Therefore, if the last complexed PPI structure, or last ligand-bound structure is removed *via* one of the filtering steps, the corresponding other structures are also removed. Structures based on homology models were removed from the dataset. Non-X-ray structures and structures with resolution < 2.5 Å were removed. Sequences were aligned using NEEDLE[86], and ligand-bound structures displaying < 80% sequence identity to their corresponding chain in the complexed PPI structure were removed. The single homoprotein complex (1TNF, representing tumor necrosis factor α) was removed, as it is both the only homoprotein complex in the dataset, and it is of questionably trimeric nature rather than binary association. The resulting 2P2I dataset consists of 16 unique PPI complexes, with 204 ligand-bound, inhibited PPI structures. This dataset is analyzed in two parts, the 2P2I P-P dataset consisting of the 16 unique PPI complexes, and the 2P2I P-L dataset consisting of the 204 ligand-bound PPI structures.

5.3.2 **File Preparation**

For the ligand-bound P-L 2P2I dataset, structures were aligned to their appropriate same-sequence partner in the complexed P-P 2P2I structure using HwRMSD[194] alignment. This was necessary to use the same calculated plane from the corresponding complexed PPI structure to analyze the ligand-bound structures, where there is no second protein. Defining a plane for the P-L 2P2I structures would not be possible, otherwise. The ligands were also removed from the P-L 2P2I structures, as they are not analyzed in this work.

The files of the PDBbind dataset required no preparation.

5.3.3 **Determining PPI Contacts and Picking Chains**

Two sets of contact atoms are made for each dataset. The first set is simply the C_α of any residue which contacts the opposite subunit in the complex, based on a 4.5 Å cutoff. The second set is specifically the exact atoms within that 4.5 Å distance of the opposing subunit.

5.3.3.1 PDBbind dataset

PPI contact residues were determined using a 4.5 Å distance cutoff *via* an in-house perl parsing script which utilizes the 3D coordinates in the PDB files. The two sequences of interested are dictated by the index file provided with the PDBbind set. This index lists which two chains the binding affinity describes. In the case where multiple copies of these sequence pairs existed in the PDB file, the two chains which resulted in the largest number of contacts were chosen as the chains to represent each structure. The Uniprot IDs for the sequences of each chain were checked to ensure that the appropriate chains were considered. For example, in a structure containing a dimer of dimers (A-B and C-D), where chains A and C are the same, and B and D are the same: if A-B resulted in 100 total contact residues and C-D resulted in 115 total contact residues, C-D was the pair chosen to represent the structure. In this example, the index file likely would have listed chain A and chain B as the chains for the corresponding affinity value.

5.3.3.2 2P2I dataset

PPI contact residues for the PPI complexes were determined using a 4.5 Å distance cutoff *via* an in-house perl parsing script which utilizes the 3D coordinates in the PDB files.

As this is a protein-centric study, contacts are determined using the complexed PPI structures of the P-P 2P2I subset. Contacts from these complexed PPI structures are then projected back onto the ligand-bound structure, since the ligand-bound structures in the 2P2I set are missing one of the subunits in the complex (as they are inhibited). Due to this, any residues which are not resolved in the ligand-bound interface structure are removed from the analysis.

5.3.4 **Plane Calculations**

5.3.4.1 The Best-fit Plane (BFP)

In-house calculation of the BFP was completed *via* 3D least-squares fitting in MATLAB[195].

5.3.4.2 SURFNET — PRINCIP

The PRINCIP program imbedded in SURFNET[124] is a direct implementation of the methodology used by Jones & Thornton[43] (discussed in section 5.2.6.2). The FORTRAN source code was altered slightly, so that the input for the software could be a full directory of PDB files, rather than one file at a time.

5.3.4.3 The Interface Geometric Centroid (IGC) Plane

To construct the IGC planes, the geometric centroid of each protein chain (each side of the interface) was first calculated as the mean of all C_α atoms in that chain's contact residues. The centroid of the entire interface was then considered to be the mean of the two chain centroids. The plane was constructed from the normal vector between the two chain centroids, by forcing the plane to intersect with the interface centroid.

5.3.5 Planarity Calculations and Protrusion/Hollow Distances

The planarity RMS (pRMS) metric, previously implemented by Jones & Thornton[43], is calculated as the root-mean-square deviation of all points in the interface from the interface plane. In their work, the best-fit plane was used. In this study, the best-fit plane was tested, but ultimately the IGC plane is used for analysis. The displacement of each atom from the plane is the same as the normal-vector distance for each atom. The normal-vector distances are solved by applying the plane's normal vector as a velocity to each atom's coordinates to calculate the necessary distance to be traveled to intersect the plane.

The distance which any atom protrudes beyond its subunit's side of the plane is aptly referred to as its protrusion distance. Conversely, the distance which any atom resides "behind" the plane on its own subunit's side is deemed the hollow distance.

5.3.6 Point Projection and Clustering

Before clustering, plane-projected atomic coordinates were obtained. The normal vector of the plane was utilized to project where any atom would intersect with the plane, in the normal-vector direction. Essentially, the three-dimensional coordinate space is flattened to two dimensions on the interface plane.

Clustering was performed using a DBSCAN clustering algorithm implemented in R[114]. Parameters for DBSCAN clustering included $\varepsilon = 3$ Å (epsilon: Euclidian distance for clustering), and minimum points per cluster of 3.

5.3.7 Protein and Plane Graphics

Graphics displaying protein structures, as well as those with graphical representations of planes were created using the PyMOL[196] molecular graphics software.

Plane drawing was heavily assisted by a publicly available plane drawing script which automates the cGO plane drawing functions of PyMOL. This script can be found at found at: http://pldserver1.biochem.queensu.ca/~rlc/work/pymol/draw_plane_cgo.py (Accessed: 7/9/2018)

5.4 Results and Discussion

The focus of this work is to quantify localized, geometric structural differences that exist between "druggable" and "less druggable" PPIs. For this context, druggable PPIs are simply defined as those which have known small molecule inhibitors, and less druggable PPIs are those which have not yet been successfully modulated by small molecules. While some arguments may be made for the differences in "bindable" and "druggable", we will address the targets simply in the "druggable" context for this work for brevity.

5.4.1 Datasets

We employ two datasets herein: a subset of the 2016 2P2I db dataset[57] (which will be referred to as the 2P2I set) and a subset of the 2016 PDBbind PPI dataset[61] (which will be referred to as the PDBbind set). The 2P2I set contains 16 unique heterodimeric PPI complexes which are represented by at least one structure of the complexed PPI subunits and at least one structure of a ligand-bound subunit where the complexation was inhibited. There are 204 total ligand-bound PPI structures in the 2P2I set. The PDBbind set consists of 347 complexed heterodimeric PPI structures for which there are K_d binding affinity data.

These two datasets share no redundant proteins and are intended to represent druggable targets (2P2I set) and less druggable targets (PDBbind set). The acquisition and filtering of these datasets, along with any necessary preparations to the files are presented in sections 5.3.1 & 5.3.2 of the methods.

Each of the datasets has relevant subsets of data. The 2P2I set can be separated into the complexed PPI structures and the ligand-bound PPI structures. The PDBbind set has three total subgroups of data, based on binding affinity. Theoretically permanent protein-protein complexes have a dissociation constants in the *nM* range or lower[169] ($K_d < 1 \times 10^{-9}$ *M* or $pK_d > 9$), while weakly transient complexes have dissociation constants in the μM range or higher[169] ($K_d > 1 \times 10^{-6}$ *M* or $pK_d < 6$). We therefore classify the complexes with dissociation constants between those ranges (1×10^{-9} *M* $< K_d <$ 1×10^{-6} *M* or $6 < pK_d < 9$) as strong transient complexes. The populations of these weak transient, strong transient, and permanent subsets are 97, 197, and 53 proteins, respectively.

5.4.2 **Defining PPI Residues**

Early approaches for defining PPI residues involves calculating SASA for separated protein subunits and then calculation of SASA for the complex.[51] Interacting residues were first defined as those that lost any amount of SASA[51]; later approaches adopted a threshold of necessary SASA loss in order to rule out small errors in calculations[43]. This adapted ΔSASA method with the small threshold (1.0 $Å^2$ for residues, 0.01 $Å^2$ for atoms) was the continued method of choice throughout the rest of the geometric studies of PPIs.[40, 43, 47, 52-54, 197]

This method is seemingly more robust than the traditional distance-based cutoff for acquiring protein-ligand contacts because it includes residues that are not in direct contact with the contrasting protein chain, but it also misses some residues that are in close proximity according to our early experiments (data not shown). It seems the primary goal of this SASA-based detection method was to embrace the relationship between solvent accessible nonpolar surface that becomes buried upon complexation and the enthalpic contribution of the hydrophobic effect as discussed in 1.2.1.[27] While this is admirable, the assumption that the tertiary structure of these protein subunits do not undergo conformational change upon complexation is required for the SASA-based method, and it is likely not always true. Even so, global changes in SASA are not qualitative on the atom-level, and do not shed light on the localized geometry of the protein surface that we aim to study in this work.

We opted to focus our methodology on characterization of specifically the protrusions and hollows present in these interfaces. Complementarity of the interface halves is unquestionably necessary to facilitate solvent evacuation, hydrophobic collapse, and thereby complex formation. Perhaps the fitting of the protrusions is responsible for some degree of this complementarity.

Preventing association of the protein complex through mechanical disruption of this crucial seating mechanism *via* modulation by small molecules is plausible.

Since the protrusion/hollow contact points are likely to represent the majority of the close-proximity atom pairings in the interface, we elected to use a distance-based cutoff to determine PPI contact residues, similar to protein-ligand binding analysis. Our adaptation of the plane-based analysis method begins with determining PPI contact residues *via* a 4.5 Å distance cutoff. Cutoffs of up to 6.5 Å were investigated but led to overly biased fitting of planes due to the exaggeration of atoms contacted by the protrusions.

5.4.2.1 Resulting PPI Contact Residues

The 2P2I dataset has an added layer of complexity when describing its data. Since we have obtained a ligand-bound complex with one of the two subunits present in each of the 16 PPI complexes, the chains in the PPI complexes can be designated as the "druggable" chain and the "complementary" chain. Note that none of the PPIs have structural data for ligands bound to both of their subunits; the ligand-bound structures all represent the same subunit. The druggable chain corresponds to that which we have obtained a ligand-bound structure for, and the complementary chains are those which are displaced by the ligands in the set. These subsets will be referred to as the 2P2I P-P druggable and 2P2I P-P complementary sets. For the ligand bound structures in the 2P2I set (2P2I P-L set), the contacts from the corresponding protein-protein complex are used to assess the structures, since one of the subunits for the overarching PPI is missing. Only the 2P2I P-P druggable and complementary chains are addressed for the analysis of the contact residues to follow, as their contact residues are used to define the contacts in the 2P2I P-L set.

For ease of presentation, the amino acids have been divided into four chemical types: nonpolar aliphatic (Ala, Gly, Ile, Leu, Met, and Val), polar uncharged (Asn, Cys, Gln, Pro, Ser, and Thr), polar charged (Arg, Asp, Glu, His, and Lys), and aromatic (Phe, Trp, and Tyr). The complete distributions of contact residues by specific amino acids may be found in Appendix B. A distribution of the resulting PPI contact residues by chemical class for the druggable and complementary chains of the 2P2I set is presented in Figure 5-2. The composition of the whole interfaces for the ligand-bound P-L 2P2I structures are not presented, as their interfaces are defined by those of the P-P 2P2I set and are thus identical to their druggable chains. The PPI contact residues of both combined chains of the weak transient, strong transient, and permanent complexes of the PDBbind set are presented by chemical type in Figure 5-3.

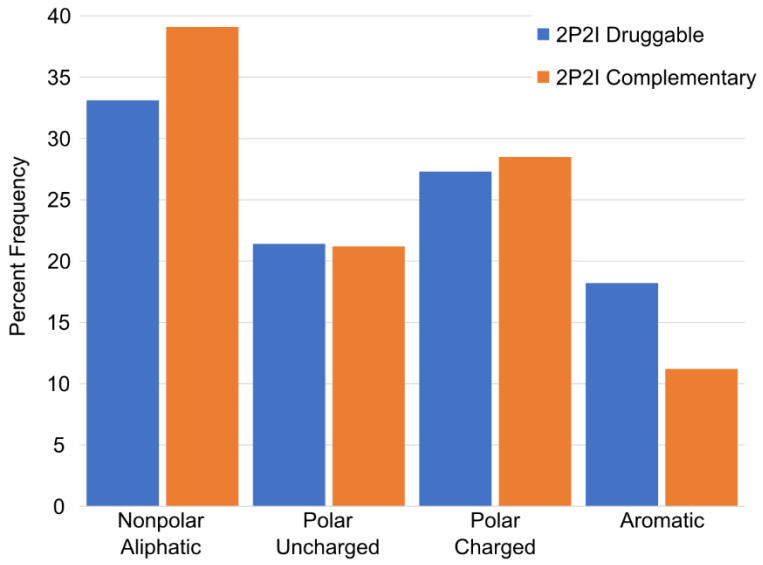


Figure 5-2. Distributions of PPI contact residues by chemical type for the 16 PPI complexes of the 2P2I set. Interface contact residues belonging to the druggable chains and complementary chains.

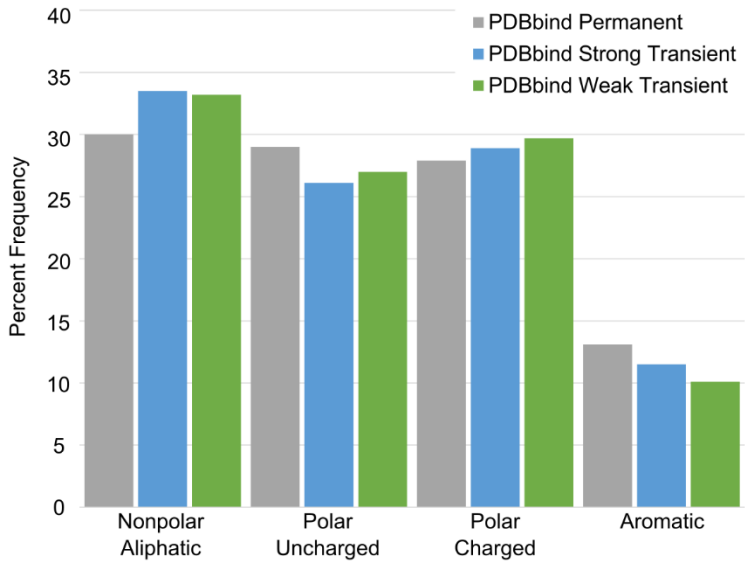


Figure 5-3. Distributions of PPI contact residues by chemical type for the complexes of the PDBbind set. Interface contact residues belonging to the permanent, strong transient, and weak transient complexes.

Interestingly, the druggable and complementary chains of the 2P2I dataset display distinct differences in composition (Figure 5-2). The general populations of polar uncharged and polar charged residues for the two sets are similar. However, druggable chains have a higher composition of aromatic residues (especially Tyr) while complementary chains have a higher composition of nonpolar aliphatic residues (especially Gly and Leu).

For the PDBbind data, both the weak and strong transient sets show nearly identical distributions of amino acid composition (Figure 5-3). The permanent complexes show a shift in composition away from the transient states, with a decrease in nonpolar aliphatic residues (especially Leu) and polar charged residues and an increase in polar uncharged and aromatic residues.

While these variances in composition are intriguing, perhaps assessing the localized topography of these interfaces would yield more pertinent information about the differences in subunit affinity and druggability. For this, we must assess the interfaces more closely.

5.4.3 **Plane Fitting**

We aim to gain information about the topographical details of these PPIs, so Thornton's method of utilizing planes and exploring deviation from those planes served as a logical place to start. The metric of planarity RMS (pRMS) is defined as the root-mean-square deviation of all relevant points (interface atoms) from the plane in question and is pertinent to quantifying the flatness and relative shape of the interface being investigated.[43] Relatively planar interfaces are defined as those exhibiting pRMS $\leq$ 6 Å. It became immediately apparent that the best-fit plane (BFP), the result from least-squares regression, can be improperly biased in two major ways.

First, the best fit to a cloud of points representing the interface atoms may not yield a plane that bisects the interface, but rather a plane that bisects the globular proteins. Numerous occasions of this event were observed during early testing of BFP analysis, where we utilized an in-house BFP calculation implemented in MATLAB[195] (see Methods, section 5.3.4). This event, which we will refer to as complex-bisection, is observed for all four of the major geometric shapes of interfaces in our datasets. It has seemingly little to do with the geometry of the interface, and it became increasingly prevalent when coupled with the ΔSASA-based definition of interaction residues. This observation is based on the results of analyzing our dataset with the PRINCIP program within the SURFNET[124] software, an implementation of Thornton's[43] analysis methods. Figure 5-4 displays the bisection of the globular protein chains in the complex of Titin with obscurin-like protein 1 (PDBid: 2WP3). This example is crucial because the interface in 2WP3 is extremely planar. The BFP as a mathematical concept was incomplete for accommodating a geometrically simple interface.

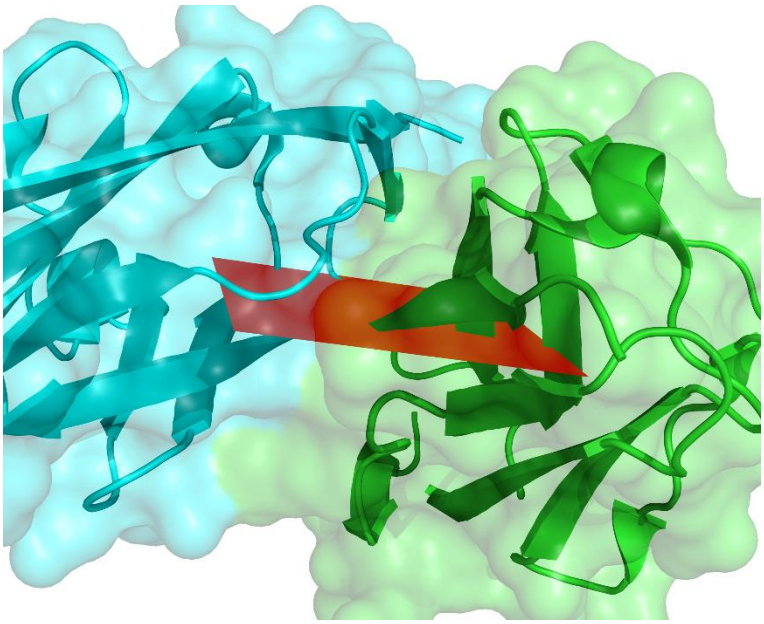


Figure 5-4. Complex-bisection of the globular protein in 2WP3 by the mathematically best-fit plane. The best-fit plane calculated using Thornton's method implemented in the PRINCIP program, part of SURFNET, fit to the planar interface of 2WP3 yielding a pRMS of 6.26 Å.

Second, the fit of the BFP is biased towards the hollows of protrusion-hollow interactions. The mathematical nature of scanning for new contacts (atoms in our case) from single reference point (a protrusion) into relatively densely populated coordinate space (the hollow in the adjacent protein chain) yields many new explored points for that single referenced exploration starting point. Due to this, protruding a single side chain or loop fragment into an opposing partner's space results in many more atoms detected on the "receiving" side of a protrusion in an interface. This is one of the very events that causes the BFP to be overfit, especially in an "all atom" context. Due to this, we aimed to define our planes in such a way that they would be more fit towards the protein backbone representation of each side of the interface and not the side chains. To accomplish this, we first tried the BFP method, fitting to only the C_a atoms of the protein-protein contact residues. These planes were deemed to still be inadequate, as the complex-bisection problem was no less frequent. At this point we elected to forego the BFP method and adopt a new approach.

Planes only require a normal vector and a single point along that vector to be defined, so adapting new approaches was merely a task of picking appropriate calculable reference points from which to create planes. Early versions of this methodology utilized two separate planes, one of each side of the interface, in attempt to quantify the relative shape of the interface by the angle created between the two interface planes. From this, we learned that the orientation of the globular

protein has little impact on the exact geometry of the created PPI, and our approach to shape quantification needed to be more localized to the exact interface. From this we defined the interface geometric centroid (IGC) plane-fitting method. Each side of the PPI is summarized using the geometric centroid of all C_α atoms for the protein-protein contact residues (which could also be deemed the center of mass, since they are all carbon atoms). The vector between these two centroids is deemed the normal vector, and the mean point between the two centroids is considered the center of the interface. As an example, we observe the fit of the IGC plane for 2WP3 in Figure 5-5, this is the same protein example used in Figure 5-4. The better representation of the IGC plane is immediately revealed, with a pRMS of 2.34 Å vs the BFP pRMS of 6.26 Å.

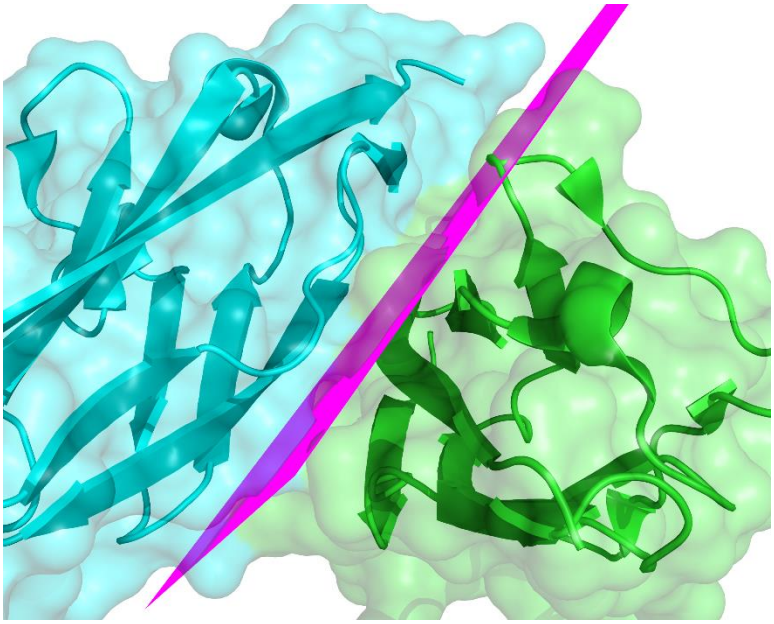


Figure 5-5. IGC plane fit to 2WP3, a planar interface with a pRMS of 2.34 Å.

The planarity pRMS was initially presented with the expectation that proteins exhibiting >6 Å of pRMS were expected to not be planar interfaces.[43] This expectation was to represent the point where non-planar interfaces would result in sets of points that would yield a plane which did not adequately represent the proteins at all. However, the median PRINCIP BFP pRMS for the PDBbind dataset is 8.01 Å, indicating that most of the entire dataset is far from planar, which cannot be true. Conversely the median IGC plane pRMS is only 4.25 Å, with 263 of the 347 PPIs displaying pRMS < 6 Å for this method. Figure 5-6 presents the distribution of pRMS values for the IGC planes and the BFPs calculated with PRINCIP. Importantly, the pRMS of the PRINCIP calculated BFPs are similar to those calculated by our in-house BFP script (data not shown), which

also yielded many planes that bisected entire complexes. For these reasons, only the IGC planes will be utilized from this point forward. The distribution pRMS values for the weak transient, strong transient, and permanent subsets of the PDBbind dataset are nearly indistinguishable, so they are provided as one set in Figure 5-4.

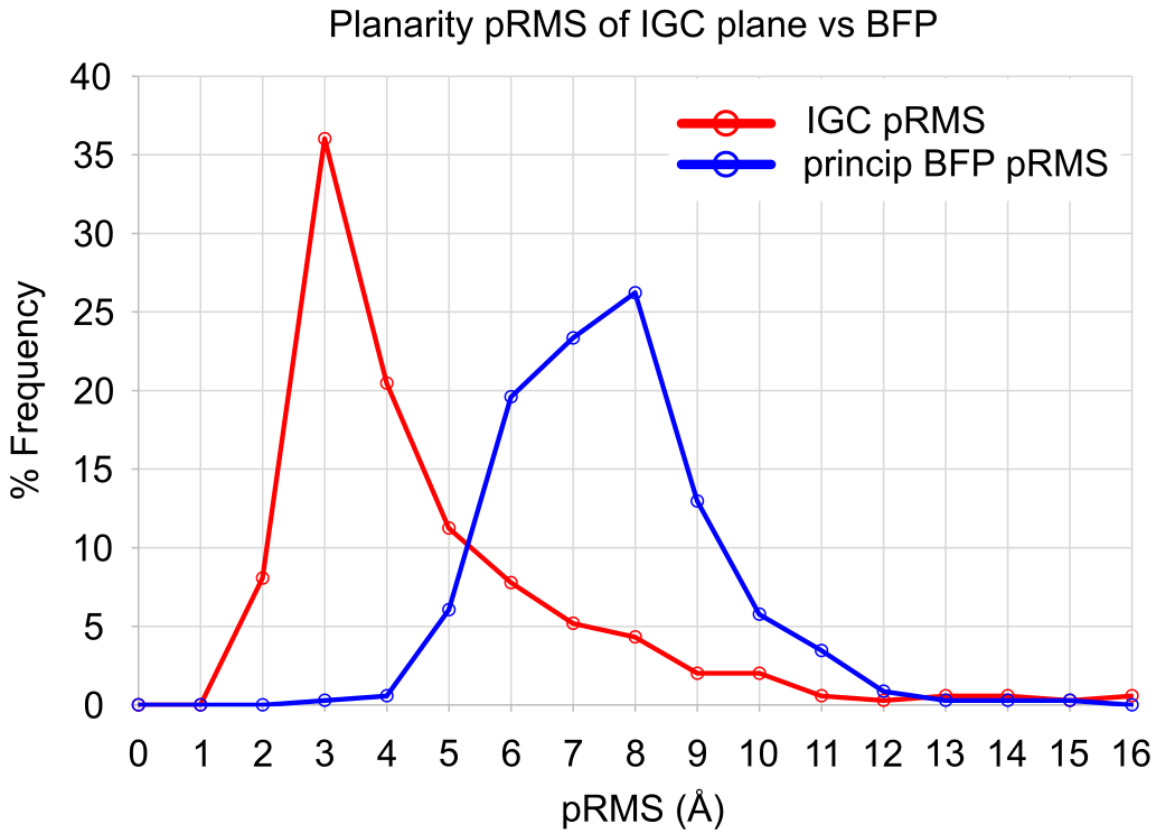


Figure 5-6. Distribution of pRMS values for IGC planes and PRINCIP best-fit planes for the 347 proteins in the PDBbind set. IGC median = 4.25 Å, PRINCIP BFP median = 8.01 Å, distribution binned *via* left endpoint (i.e. pRMS of 2 displays 2.0-2.99)

In order to learn about the interface landscape of the P-L structures in the 2P2I dataset, aligning of the structures was necessary. As these structures do not have both subunits of the PPI, planes cannot be calculated for them. Instead, the ligand-bound subunits were aligned to the appropriate partner in the corresponding complex structure of the P-P 2P2I set using HwRMSD[194] for the best alignment. The pRMS values of the IGC planes for both the P-P and P-L 2P2I sets are provided in Figure 5-7. The pRMS values of the P-L structures are slightly inflated, which is to be expected because they only represent one side of the total interface that the planes are calculated from. Even so, 94% of their pRMS values are < 6.0 Å which indicates that the alignment worked

appropriately and no extreme conformational changes are observed in the interfaces of the ligand-bound structures.

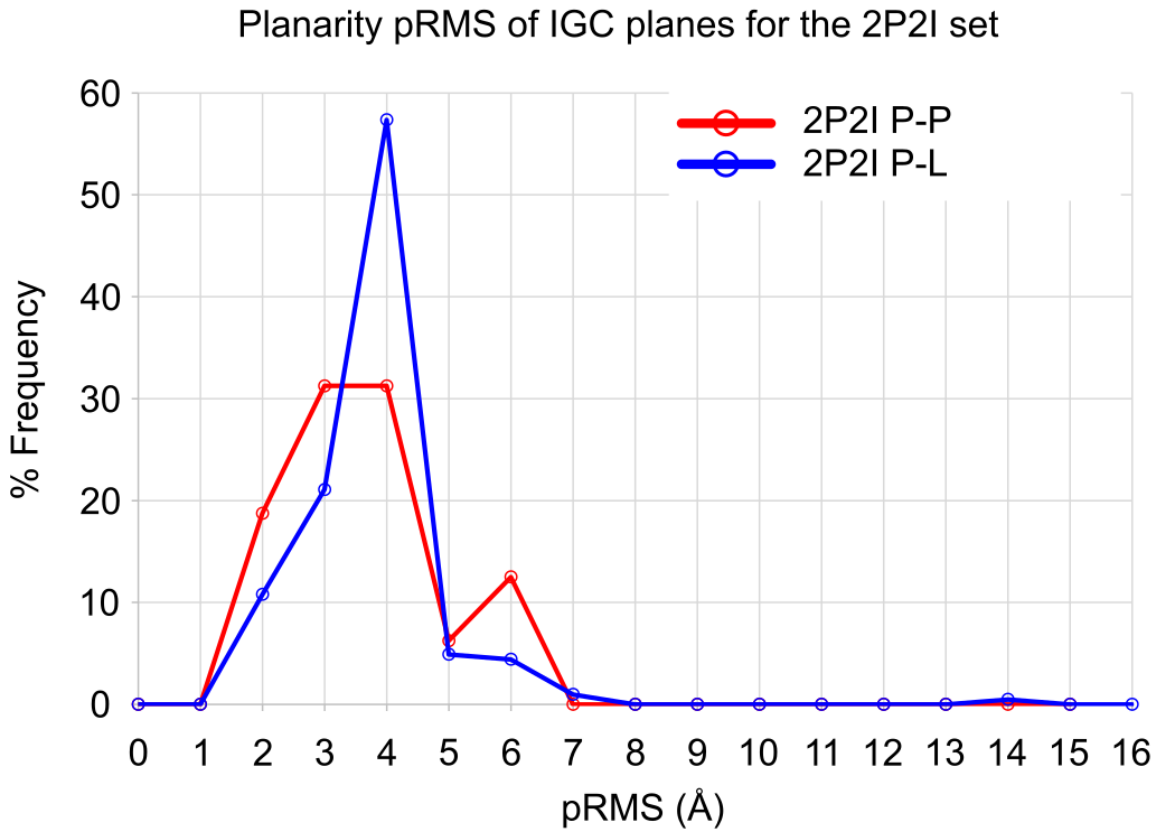


Figure 5-7. Distribution of pRMS values for IGC planes for both subsets of the 2P2I dataset.

5.4.4 Local Structure and Clustering

To discern information related to the topography of the interface residues, the displacement of the interface atoms was calculated using the normal vector from IGC plane (see methods, section 5.3.5). These are the same distances that were used for calculation of the planarity pRMS. The atoms for each subunit of the interface are then separated into two groups: protrusions and hollows referring to those points which protrude past the plane and those that reside 'under' the plane with the bulk of their protein chain, respectively. To reiterate, each interface now has four sets of data, for example: chain A protrusions (which reside more towards chain B, on the opposite side of the plane), chain A hollows (residing on chain A's side of the plane), as well as the converse groups of chain B protrusions, and chain B hollows.

Earlier work employed the use of 'segmentation', or binning the interface residues into groups based on the residue number (resn) in the protein sequence. As a pseudo clustering method, the Δresn served as a Euclidean distance, where residues were required to be within Δresn < 6 of any residue present in a segment to be added to that segment. While this approach was functional in practice, it is not always relevant. For instance, in a case where part of an interface consists of two adjacent beta sheets, residues that are directly interacting with each other could be sorted into separate segments purely because of their distance from each other in the sequence. Many details of the interfaces in past studies were undoubtedly missed by that segmentation clustering. To improve upon this, a modernized clustering approach was taken. Two crucial details were noted for success of the chosen clustering algorithm: the number of clusters and the shape of the clusters in the interface were not known. A DBSCAN clustering algorithm appealed as the best approach.

The aim of clustering was to characterize the individual protrusions and hollows that made up the interface. However, because we only had the contacting atoms of the residues, and not the entirety of the residues, mathematical compensation for the arbitrary distances between the atoms was necessary. The four groups of interface atoms were projected onto the plane, again utilizing the normal vector (section 5.3.6 of methods). Since the distances of the atoms from the interface were already known, that information was preserved elsewhere, even with the two-dimensionally flattened data. With the relevant interface atoms compressed into 2D space, they were clustered using a DBSCAN algorithm implemented in R[114]: requiring a minimum of 3 neighbors (ε) and using a Euclidean distance of 3.0 Å (section 5.3.6 of methods).

5.4.4.1 Physical Characteristics of Clusters

The four sets of grouped atoms logically form two sets of complementary features: one subunit's protrusions should seat into the adjacent subunit's hollows and vice versa. This variety of analysis is best suited for the 2P2I dataset, where the druggable subunit is known. The hollows found in the druggable subunit represent pockets where small molecules have potential to bind. Nonetheless, this analysis was applied to the PDBbind dataset in attempt to discern any possible differences in the behavior for the more strongly complexed subunits.

In the pursuit of determining whether PPIs are truly "flat and featureless" [41, 44-46], we begin with the former half of "Are they flat?". The physical characteristics of the interfaces and clusters will be assessed using the number of clusters, depth of those clusters (D_{max}), and the shape of those

clusters. We define the shape of the clusters *via* a shape index (SI), where $SI = \frac{D_{max}}{D_{RMS}}$, where D_{max} is the maximum depth of any single atom that resides in that cluster and D_{RMS} is the root-mean-square depth of all atoms belonging to that cluster. As the SI approaches 1, the cluster shape becomes "dull" in that the maximum depth of the cluster is very close to all of the other depths in the cluster, resulting in a shape like a wide mesa. As the SI increases, the cluster's shape resembles more of a pointed object, such as a typical stalactite or icicle. Table 5-1 presents the medians for number of clusters, D_{max}, and SI for protrusions and hollows in all data subsets.

Table 5-1. Physical characteristics of the interface clusters for all data subsets.
Reported errors are standard deviation.

	Median # Protrusion Clusters	**Median # Hollow Clusters**	**Median Protrusion D_{max}**	**Median Hollow D_{max}**	**Median Protrusion SI**	**Median Hollow SI**
2P2I P-P Druggable	**2 ± 1.5**	**2.5 ± 1.8**	5.4 ± 2.9 Å	5.7 ± 2.6 Å	1.53 ± 0.27	**1.56 ± 0.31**
2P2I P-P Complementary	**1 ± 0.9**	**2 ± 0.9**	4.1 ± 2.3 Å	7.6 ± 3.4 Å	**1.61 ± 0.24**	1.49 ± 0.27
2P2I P-L	**3 ± 1.1**	**2 ± 1.7**	5.7 ± 2.9 Å	5.3 ± 3.0 Å	1.46 ± 0.30	1.58 ±0.28
PDBbind Permanent	5 ± 2.1	8 ± 3.6	4.3 ± 3.7 Å	6.6 ± 3.8 Å	1.52 ± 0.27	1.46 ± 0.34
PDBbind Strong Transient	6 ± 2.9	9 ± 3.4	4.1 ± 4.6 Å	5.8 ± 4.5 Å	1.51 ± 0.28	1.47 ± 0.31
PDBbind Weak Transient	5 ± 2.0	8 ± 3.0	3.8 ± 3.5 Å	5.2 ± 3.8 Å	1.53 ± 0.27	1.47 ± 0.31

Bolded table values indicate values which are explicitly discussed in the text.

The abundance of protein-peptide complexes and protein-protein complexes with very small contact areas in the 2P2I set becomes immediately apparent, as represented by the smaller number of clusters in the different 2P2I sets. That aside, the characteristics of their protrusions and hollows are very similar to those of the PDBbind sets. The most crucial details are the median protrusion SI for the complementary chains in the P-P 2P2I set coupled with the median hollow SI for the druggable chains in the P-P 2P2I set, as well as the median hollow SI for the 2P2I P-L structures. These measurements suggest that the hollows in the "druggable" subunit of the PPIs are shaped more sharply than other hollows found in the interfaces. These druggable hollows are not necessarily any deeper than the hollows in the rest of the interface, but a sharper shape would allow bound ligands to be less solvent exposed.

The measured parameters for the three affinity-based PDBbind subsets are extremely similar. All three affinity-types of the PDBbind set have more hollows than protrusions, and hollows that are deeper than their protrusions reach. Perhaps these extra hollows are filled by very

small protrusions of residues. Our clustering approach required 3 nearby neighbors for clusters to be kept, which results in 97% of the structures having at least one un-clustered data point in their hollows and 95% of the structures having at least one un-clustered data point in their protrusions.

Another metric of flatness could be described as the total deviation of the interface surface about the plane, akin to an 'elevation' change relative to the plane. This change in elevation could be measured as: $\left|D_{max}^{Protrusion}\right| + |D_{max}^{Hollow}|$. As a conservative estimate, the lowest D_{max} observed in the protrusions of any group is ~3.8 Å (weak transient PDBbind set), and the lowest D_{max} observed in the hollows of any group is ~5.2 Å (weak transient PDBbind set). These values would yield a net 9 Å of maximal elevation change relative to the plane across the interfaces of those proteins. While this metric does not describe the gradient of elevation change across the protein surface or account for the relative size of the interfaces, we can safely conclude that these interfaces are surely not flat.

5.4.4.2 Chemical Characteristics of Clusters

The assessment of these PPI targets continues with investigating chemical features present in these protrusions and hollows. The most characteristic residue of a protrusion or hollow is likely the deepest residue which is contacted by the opposing subunit. These residues were represented as the deepest atom in the cluster from which the cluster depth (D_{max}) was calculated. Here, the residue composition of the interfaces are presented once more, but only for the representative residue of each cluster. For clarity, these representative residues are those from which the largest depth (D_{max}) was calculated from the plane, either past the plane towards the adjacent subunit (protrusions) or 'behind' the plane towards the subunit to which the residue belongs (hollow). The distributions of representative protrusion and hollow residues for the druggable and complementary chains of the P-P 2P2I dataset are presented by chemical type in Figure 5-8. The order that the data is presented in Figure 5-8 has significance, as it portrays the nature of the data groups in the actual PPI, i.e. the druggable protrusions fit into the complementary hollows. The complete distribution by individual amino acid types for the P-P 2P2I set can be found in Appendix B.

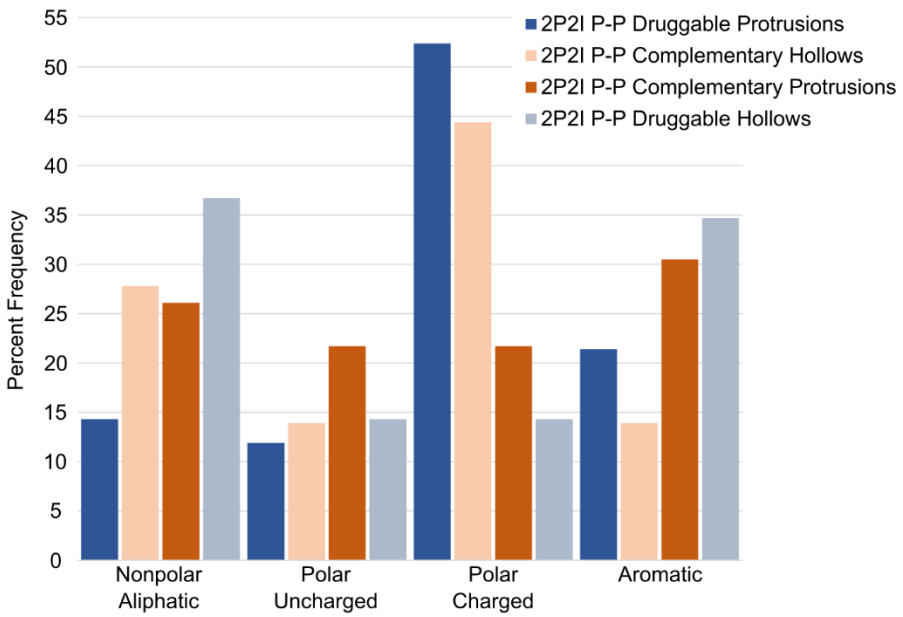


Figure 5-8. Distributions of PPI contact residues representing the protrusions and hollows for the 16 PPI complexes of the P-P 2P2I set by chemical type. Representative cluster residues belonging to the druggable chain protrusions, complementary chain hollows, complementary chain protrusions, and druggable chain hollows.

It is immediately evident that the druggable chains primarily contribute protrusions of polar charged nature with complementary chains reciprocating that overwhelmingly polar charged nature in their hollows. Conversely, the complementary chains contribute primarily nonpolar aliphatic, as well as aromatic protrusions which are complemented by similar hollows in the druggable chain. These interaction types follow a logical cascade of events for PPI complex formation. The hydrophobic collapse of the specifically nonpolar and aromatic protrusions and hollows function to bring the polar charged protrusions and hollows into close proximity so that salt bridges may be formed, locking the interface in its complexed state. However, upon small molecule binding to one of these generally nonpolar pockets, either aliphatic or aromatic in nature, closing the gap of hydrophobic collapse becomes too difficult. The result of this small-molecule block prevents the appropriate polar charged moieties of the two subunits to ever reach close enough proximity to solidify the interface. Thus, the formation of the interface is inhibited. The role of this chemical complementarity between the two subunits cannot be coincidental.

Figure 5-9 presents the distributions of representative cluster residues for the protrusions and hollows of the 204 ligand-bound P-L structures of the 2P2I set, alongside the druggable chains of the corresponding P-P structures. These two sets of data represent the same functional subunit

of the PPI and their chemical characteristics should be similar, so they are presented as a pair. The representative cluster residues for the ligand-bound structures are important, as any differences when compared to the druggable chains from the complexed P-P structure could indicate conformational changes due to ligand binding.

The protrusions of the ligand-bound subunits show an extreme abundance of charged polar residues, while their hollows are comprised primarily of aliphatic nonpolar and aromatic residues. Their distribution of residue types mimics that of the druggable chains in the complexed PPIs which they represent, with the ligand-bound distribution being even further skewed. It is possible that the protrusions of the ligand-bound conformations display an increased representation of charged polar residues because those protrusions are solvent exposed in the ligand-bound structures. Similarly, when compared to the druggable chains of their complexed counterparts, the hollows of the ligand-bound structures show a slight increase in uncharged polar residues. This could represent small conformational changes by the subunit in attempt to hide its hydrophobic core residues from the interface surface, which is now mostly solvent exposed due to inhibition of complex formation.

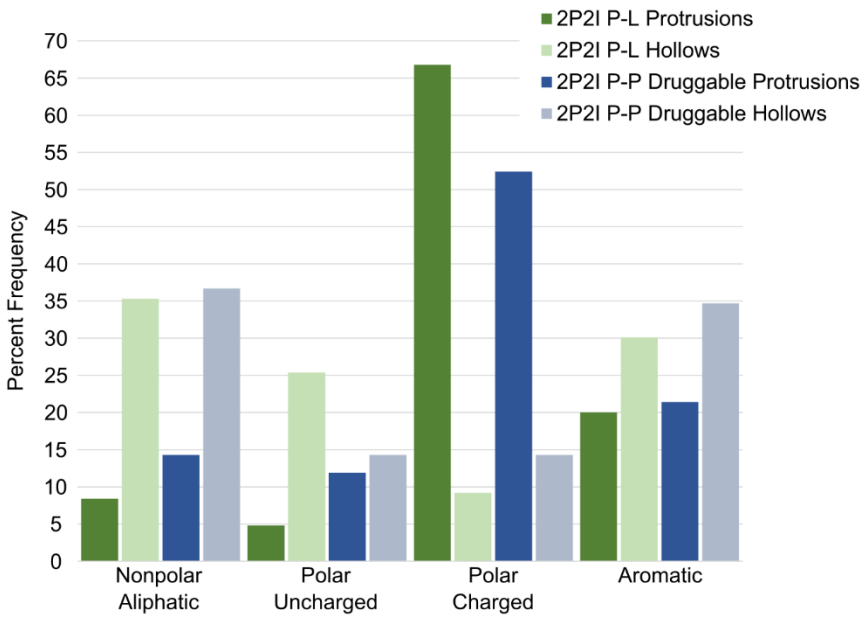


Figure 5-9. Distributions of PPI contact residues representing the protrusions and hollows for the ligand-bound PPI complexes of the P-L 2P2I set and the druggable chains of the P-P 2P2I set. Representative cluster residues belonging to the ligand-bound chain protrusions and hollows, presented alongside the protrusions and hollows of the druggable chains of their complexed PPI counterparts.

Finally, Figure 5-10 presents the distribution of the representative residues for the protrusions and hollows of the permanent, strong transient, and weak transient complexes of the PDBbind set, separated by chemical type. While the permanent complexes show a slight bias towards more aromatic residues, in both their protrusions and hollows, all groups of PDBbind structures show very similar distributions of side-chain functionalities.

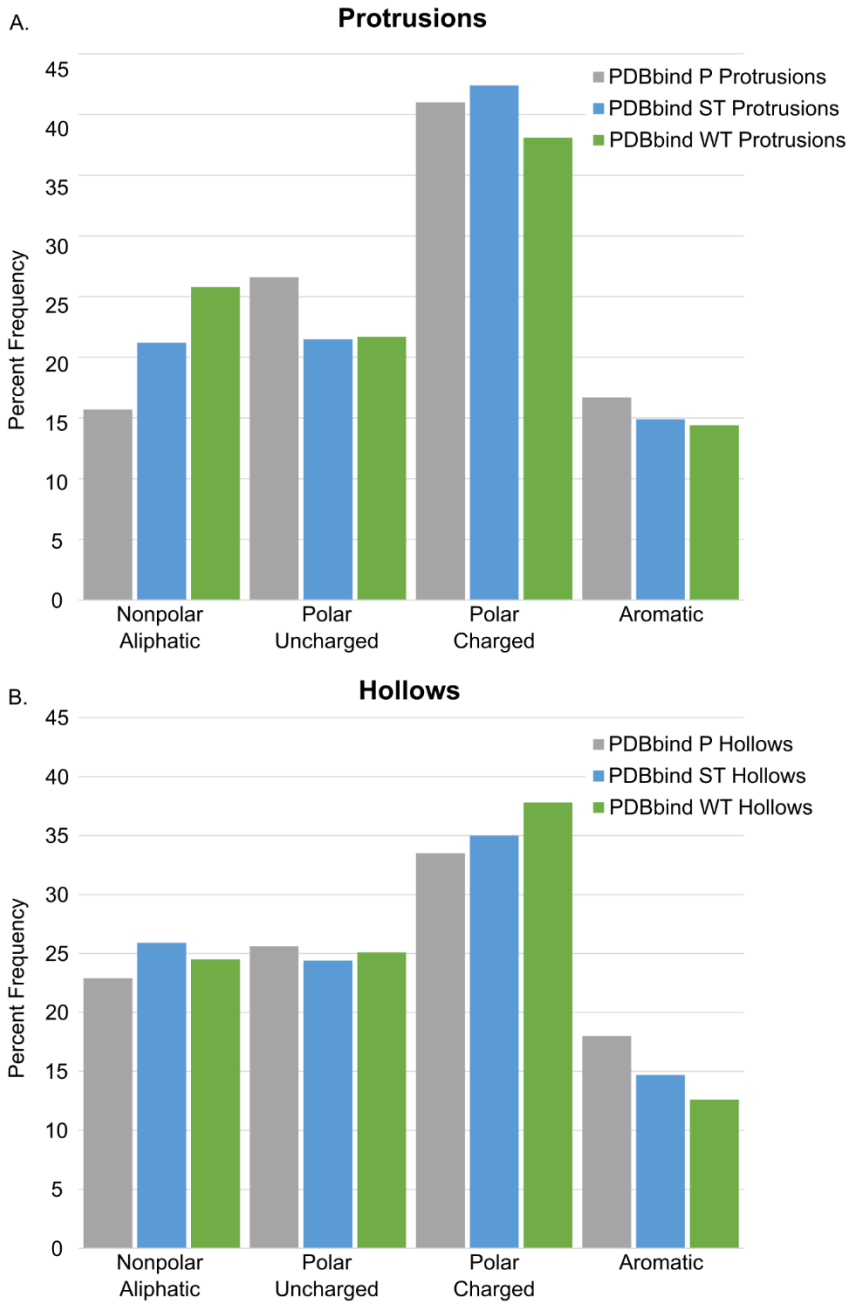


Figure 5-10. Distributions of PPI contact residues representing the protrusions and hollows for the complexes of the PDBbind set, by chemical type. Contact residues belonging to the: permanent, strong transient, and weak transient subsets of complexes separated by their A) protrusions and B) hollows.

5.4.5 Relationships to Other Metrics

The primary purpose of seeking a less druggable set of data with the added element of binding affinity for the complex formation was to assess whether any relationships exist between these affinities and other physical parameters of the interfaces. However, attempting to draw correlations between binding affinity and interface size (# res), ratio of interface sizes (larger chain # res/smaller chain # res), pRMS, number of clusters, cluster depth, and cluster shape index all resulted in R^2 values < 0.1.

Nooren & Thornton concluded that the $\Delta G_{binding}$ between protomers was not related to the size, planarity, or polarity of the interfaces that they investigated.[40] Finding that we could not distinguish between the affinity-based subsets of less druggable complexes using our cluster shape and depth parameters, or the chemical makeup of those clusters is therefore disappointing, but not unexpected.

5.5 Conclusion

We investigated the physical shape and chemical makeup of the protrusions and hollows found in the interfaces of the druggable PPIs represented by the 2P2I dataset, as well as the theoretically less druggable targets of the PDBbind set. We have established a new protocol for generating planes to represent the PPI interfaces, the interface geometric centroid (IGC) method. This new method displayed better planarity scores than the best-fit plane (BFP) method used in the past when compared both to our in-house implementation of the BFP method and to BFPs calculated by the PRINCIP program within SURFNET[124]. Not only did the IGC plane method show quantitative improvements, but it also circumvented a large issue with the BFP method in the form of complex bisection. The best-fit plane was often oriented such that it completely bisected the globular protein subunits rather than the interface between the subunits.

The PPIs represented by the 2P2I set are largely of protein-peptide nature, which is evident throughout the analysis of the clusters. The 2P2I sets had fewer protrusions and hollows than the proteins in the PDBbind set. Druggable PPIs consisting of smaller interaction interfaces is not

coincidental, but likely a representation of the limitations in inhibiting the massively coordinated complexes of PPIs with small molecules.

The respectively polar and nonpolar nature of the clustered protrusions and hollows observed in the PPI structures illuminates some necessary features of druggable interfaces. Those interfaces in which one subunit primarily contributes polar interactions and the other subunit primarily contributes nonpolar interactions may be a template for the discovery of druggable PPIs in the future. At the least, the hollows of the druggable subunit consisting primarily of nonpolar aliphatic and aromatic residues is significant. The sharply-shaped nature of the hollows in the druggable subunits, which would allow for bound ligands to better bury themselves from bulk solvent, also reiterates that buried surface area and ligand efficiency are important.

The binding affinity describing the complexation of the subunits forming a PPI had very few apparent relationships with any of the characteristics we measured, confirming the findings of earlier work[43, 47]. Pointedly, the tighter binding complexes had slightly deeper hollows and farther-reaching protrusions than the weaker complexes. Perhaps this hints at subunits with higher affinity being more interdigitated with one another. If this were true, these complexes would likely form soon after synthesis and be extraordinarily difficult to dissociate, much like homomeric complexes.

We then return to the starting quandary of "Are PPIs truly flat and featureless?". On the debate of flatness, there was no correlation between the number of protrusions or hollows and the pRMS of the interface (the planarity or flatness). This suggests that even our most planar interfaces showed similar numbers of protrusions and hollows to the more complicated interfaces. Furthermore, we showed that even the most "flat" interfaces in terms of deviation about the plane had a median of 9 Å of change in 'elevation' relative to the plane. While this value does not describe the gradient at which the surface changes elevation, or account for the size of the total interface surface, it demonstrates that these surfaces are not flat.

The presence of many hollows offers an abundance of small pockets for which ligands could be designed, but many of the hollows are likely incapable of modulating the entire PPI. With the organized nature of the chemical contributions of the protrusions and hollows in the druggable PPIs studied here, it seems that the mission may not be to find the right pocket, but perhaps the correctly ordered PPI.

Chapter 6. Conclusions and Future Directions

6.1 Significant contributions of this thesis

The introduction of this thesis (Chapter 1) provided a brief overview of relevant physicochemical properties and concepts which are referenced and utilized throughout the rest of the work. The properties of solvation events, Van der Waals interactions, and electrostatic interactions are heavily discussed. These properties are the foundation to understanding biochemical binding events and describing specific interactions of binding events in proteins. The details of small molecule binding are then contrasted to protein-protein binding events, and some relevant concepts for protein-protein interactions are presented. This chapter concludes with briefly introducing some crucial databases which are used in the experimental work of later chapters.

Chapter 2 thoroughly introduces Binding MOAD[3], a collection of high-quality, hand-curated ligand-bound crystal structures maintained by the Carlson lab. We currently have 25,759 protein-ligand complexes which represent 7599 families of proteins (when binned by 90% sequence identity) and 12,432 unique ligands. For these ligands, we have 9138 binding data, which covers 35.5% of the complexes in our dataset. My primary contribution to the latest developments in Binding MOAD is that of the data pipeline used to create the unified binding sites. Numerous formatting and protein numbering issues had to be rectified in order for assembly of these unified binding sites to be possible. As such, the inclusion of family binding site residues into the protein viewing tools available to our users are one of the newest features. We expect to complete a very large dataset update, for all new PDB structures deposited between 2015-2017, by the end of this year (2018).

Chapter 1 presents a robust study of protein flexibility upon ligand binding. The work begins with curation of a large dataset of ligand-bound (holo) and ligand-free (apo) structures for

which each unique protein sequence has at least two structures of each type. This dataset is the largest of its kind and is utilized in the experimental work of both Chapter 1 and Chapter 4. Instead of defining the binding site on a by-structure basis where each holo structure has its own binding site definition, we utilized unified binding sites. The unified binding site for each protein target represents the union of all protein residues which contact *any* ligand across *all* of the bound structures representing that target. This robust definition helps us achieve information about the extended areas of the binding site, which is especially important in larger targets.

We find that the inherent backbone flexibility across the apo structures is roughly the same as the variation across holo structures. The induced backbone flexibility across apo-holo pairs is larger than that of the apo or holo states individually, but the increase in RMSD is less than 0.5 Å. Analysis of χ_1 angles revealed a distinctly different pattern with significant influences seen for ligand binding on side-chain conformations in the binding site. Within the apo and holo states themselves, the variation of the χ_1 angles is the same. However, the data combining both apo and holo states show significant displacements. Upon ligand binding, χ_1 angles are pushed to new orientations outside the range seen in the apo states. The side-chain flexibility of each amino acid was relatively quantified in this study. Rather than rotamer libraries, actual flexibility profiles for use in flexible side-chain docking of small molecules could be derived from these results. Finally, correlations between binding site variation and features such as ligand size and X-ray structure resolution were probed, but no such relationships were found. Combining all of the flexibility-related details, we find that binding site flexibility is compatible with the common practice in flexible docking, where backbones are kept rigid and side chains are allowed some degree of flexibility.

Chapter 4 presents a benchmarking exercise of six binding-site prediction methods, where we investigate their relative performance on apo structures and holo structures. This work uses a culled down version of the dataset from Chapter 1. Most binding-site prediction benchmarks only utilize holo structures, because it is commonly believed that the methods are more successful with them. Our testing of Surfnet, Ghecom, $\text{Ligsite}_{\text{csc}}$, Fpocket, Depth, and AutoSite revealed that there were no functional differences between the performance of apo and holo structures in all methods except for Fpocket. Even in Fpocket's case, the preference for holo structures was only minute. Furthermore, for almost every protein and every method tested, there was at least one of each type

of structure where the method failed to predict any part of the correct binding site. We used the unified binding sites as our definitions in this work, which is gratuitous towards having methods not completely fail, as there are even more possible "correct" answers in the binding site.

At this point, the focus of the study shifted towards attempting to qualitatively assess why so many structures were failing, and where they were failing. We determined that there was no correlation between indicators of X-ray structure quality, such as resolution or Cruickshank DPI which is calculated from a number of structure factors, and the performance of the structures in any of the six methods. Furthermore, there was no correlation between the performance of any pair of the six binding-site prediction methods on specific structures. We expected much higher consistency across varying protein structures of the same sequence. These results have huge implications for the benchmarking of binding-site prediction methods for the entire community, as these methods appeared to succeed and fail on high-quality structures of the same protein sequence in both ligand-bound and ligand-free cases. We have demonstrated that there are very few methods which appear to show preference for apo vs holo protein structures. In order to extend this idea to other computational methodology, more high-quality datasets need to be made available to the community which have proper representation of apo structures.

Chapter 5 presents an investigation of the topographic nature of protein-protein interactions (PPIs). The physical and chemical properties of these valuable but elusive targets are assessed on a geometric level, to characterize the shape of the interfaces as well as the locations of important functionalities in attempt to decipher what makes PPIs so difficult to modulate. We aimed to address the classical sentiment of PPIs being "flat and featureless."

Investigating the druggable interfaces of 16 proteins gave light to two major points. First, these druggable interfaces contain structural hollows which are more sharply shaped than those of the less druggable targets in the contrasting dataset. Second, the amino acid composition of the protrusions and hollows of the druggable chain were markedly different than the rest of the subunits we investigated. These druggable targets specifically had hydrophobic and aromatic hollows, while having abundantly charged polar protrusions. Their partner subunits for the total PPI had the opposite characteristics. With such a high population of charged polar residues, it is likely that most of the free energy of binding for the two subunits is stored within their inevitably formed salt bridges. From this we derived a logical binding model. During the standardized

hydrophobic collapse necessary to form a PPI, the presence of a bound small molecule in one of these sharply shaped hollows of the druggable subunit would prevent the sharply shaped, complementary protrusion that is supposed to seat there. Due to this, the chains would not be able to reach close enough proximity for the necessary salt bridges to form with the charged polar protrusions and their complementary hollows.

With the measured physical parameters of the interface, we quantified a relative 'elevation' change of the interface surface relative to the mathematical plane we fit and used for the analysis of each protein. In the 'flattest' group of proteins, the weak transient subset of the PDBbind set, the median elevation change of a single subunit is on the order of 9 Å. While that distance is my no means grandiose for the context of an entire subunit, on the scale of druggable small molecules it is far from flat.

Our less druggable PPI set was chosen because all of the targets were annotated with dissociation constants describing the complexation of the protein subunits. We attempted to draw correlations between these dissociation constants and any of the physical characteristics we measured. Unfortunately, no correlations were found. We agree with previously accomplished work that any relationships between the free energy of binding between two protein subunits and measurable physical parameters of the PPI seem to be elusive.

We concluded with a few parting notes. If druggable interfaces need to have overly biased distributions of polar and nonpolar characteristics for their protrusions and hollows, with a complementary chain showing the opposite bias of chemical complementarity, druggable interfaces are likely to be rare. Perhaps pursuit of small-molecule modulation of PPIs is not a pursuit of finding the right molecules, but rather the right interface.

6.2 Future Directions

Structural information of protein targets is invaluable as a tool for the scientific community. Many characteristics of proteins have been better understood due to our ability to mathematically quantify and probe various features of protein structure. However, many details are still likely unknown to us. The information we obtained about both inherent protein flexibility and that which is induced by ligand binding is crucial, but only the first step. Unfortunately, biophysical experiments are subject to is the presence of bias in the datasets used, as well as assumptions

necessary to either perform calculations or form a sizable dataset to begin with. Future studies of protein flexibility could be improved in two major ways.

First, instead of simply assessing by amino acid type, the study could be based on the footprints of X-Y-Z tripeptide sequences, where Y is the residue being studied in the sequence. A study accomplished this way could draw parallels between the classical Ramachandran φ and ψ angles with the χ_1 angles we studied. Changes in side-chain behavior could be linked to presence of neighboring residues in sequence, as well. There undoubtedly exist strange side-chain conformations that are a product of the neighboring residue in sequence being an Arg or Lys which is stretching to form a salt bridge with another part of the protein sequence. Events like this which provide such large amount of energetic potential can surely influence other parts of the protein sequence. Without studying residues with the neighboring sequence in context, this information is lost.

The second idea for improving the results of this experiment come from an even more diverse dataset. What could be learned about the biases of the methodology used for structural characterization if we had at least two ligand-bound and two ligand-free structures that were X-ray structures, and another full set that were NMR structures for each protein? Related to this, even within one method, what bias does the space group of the crystal structure impose on the conformation of the proteins we study? A consequence of using previously established methods to grow new protein crystals of a previously studied protein is that most of the resulting structures have the same space group. Scientists are taught very early to never 'reinvent the wheel,' but are there consequences for never doing so? In the case of structural biology, I believe there may be.

Similarly, the effects of space group could be tested on the success of binding-site prediction algorithms. The results we observed for the seemingly random nature of some structures working and some structures failing, regardless of their ligand bound state, is likely a case of bias introduced by training the methods on given data. The impact of space group and other crystallographic factors are not always heavily discussed but they are important in these prediction-based contexts.

Appendices

Appendix A. Supplemental Information for Protein Flexibility

Table A-1. Index of Protein Flexibility Dataset

Family	Type	comma separated list of PDBids
1	apo	1aki,4owc,1ljh,1ps5,1uig,2f30,1jit,2hu1,4h90,3agg,4etb,4h9e,2xbs,5lym,1rcm,2cgi,4b4j,3kam,4et9,4uwn,4ngi,1lz9,3wum,1f0w,2cds,3ijv,2ybn,3a92,4hv1,2lym,1lsf,4uwu,3n9a,4lfp,2x0a,4e3u,1lzt,1lsa,4lyz,3a90,2ybm,8lyz,4b4i,4b0d,6lyz,4lzt,3a94,1yl0,1z55,1lz8,2w1x,1jj1,2bly,4h92,1n4f,4eta,4ngy,1jis,1uco,5lyt,2d4j,4h9f,4dd0,1ykx,2aub,3a95,4h9c,3p66,2a7f,4b49,1rfp,1ljf,2epe,4neb,1hsx,1lje,4owe,4lyo,1vat,4p2e,3lyt,1vdt,4nwh,3iju,1f10,4ny5,4dt3,4htq,4j1a,1qtk,3lyz,2hu3,3az6,4ngv,1yky,2ybl,2ybj,1hsw,1h87,3n9c,1hf4,4h8y,4d9z,1jpo,2g4q,3az5,2ybi,6lyt,1xek,1lcn,4ngk,2w1l,3p65,2yvb,2lyz,4etc,3wua,1ykz,3p4z,4dd2,1vdp,2d6b,1iee,4lfx,1jj3,3wxu,1lsc,4oot,2fbb,1azf,4lyt,4ng8,4ngj,1vau,3wul,4ngw,4htk,3ru5,2w1m,4a7d,4b1a,4ete,1wtm,1gwd,2ybh,1lsb,4i8s,3wmk,3zek,3a93,193l,2xjw,1bwh,3lzt,3p64,4h8z,2blx,3wpj,1hc0,3wu7,1ljk,2w1y,1hel,3rt5,4h9i,1v7s,4o34,1lyo,2c8p,3az4,1bwi,2a7d,1bvx,2d4k,2lzt,3wxt,1lys,2vb1,3txb,1jiy,1w6z,3p68,1ljg,3wl2,1qio,2bpu,4eof,2z12,1dpw,4n5r,2f4a,3m3u,4tws,194l,3agh,4h9b,5lyz,2f2n,1lse,1lks,3rnx,3a8z,2htx,2c8o,1wtn,4lgk,1ljj,4qeq,3lyo,4uwv,3wu8,4h8x,3wpk,4ow9,1lj4,4h9a,4nfv,2g4p,1dpx,3a96,1v7t,4axt,4ng1,3rz4,2zq4,4ngl,1xej,3txd,1b2k,1bgi,2z19,7lyz,2d4i,1vdq,3az7,1lza,4b4e,1lkr,1xei,1lyz,4etd,4ngo,3aw7,3exd,3wu9,3wpl,2zq3,1lji,2lyo,1ved,2d91,4htn,1lsd,4h94,3n9e,1lj3,4h9h,4h1p,4ngz,4iat,2xbr,4h91,1vds,3atn,1c10,4bs7,1lma,4h93,3lym,4hv2,1bwj,3a91,3rw8,4et8,4nhi,1lpi,4j1b,3aw6,4lym,4qy9
1	holo	4hp0,1hew,3qe8,1lzb,4tun,1t3p,1lzc,4hpi,3txj,3qng,1b0d,1uih

2 apo

1ca3,1xev,1tbt,1te3,1tb0,3d93,3koi,1teq,1hca,2vvb,3kok,2vva,3ks3,4ca2,3kon,3mwo,1xeg,2ax2,3d92,3m1j,1t9n,1teu,3kwa,2ili,1f2w,1rzd,1rzb,3gz0,2ca2,1rza,4cac,1cay,1cah,1rze,5cac,1raz,1bv3,1rzc,1ca2,1ray,3tmj

2 holo

3k34,4ht0,3m5e,3p5a,3mhc,3m96,2q1q,3m2n,3oys,3s8x,3k7k,3bl1,3mho,4n16,3k2f,3myq,4q6d,1eou,3efi,3n0n,2o4z,3t84,4knj,3n4b,4mlt,3hkn,2geh,3m3x,3b4f,4e3g,4r5b,3n3j,3caj,3ni5,3sbh,3d9z,3d8w,3m67,2fmg,3mhm,2pow,4e4a,2q1b,1cnx,2h15,3t82,1kwr,3p58,3bl0,3sax,3mhl,3p5l,2pou,3nb5,3dd8,1cny,3mmf,3sap,4q6e,3ffp,3oy0,2ez7,3mzc,3c7p,3daz,2hkk,3f8e,3dd0,3ml2,2eu3,3eft,4mlx,2fmz,3f4x,4mo8,3n2p,3m40,3sbi,2eu2,4cq0,3mna,1cnw,4lp6,3s9t,3p4v,3oyq,3t85,4e3h,4kni,1if8,3vbd,2aw1,1bnv,3ca2,3ryz,2weo,3v7x,3ibu,1bnt,2wej,1bn4,2x7t,3t5z,3ibl,1if6,2wd2,1okm,3rz5,1cim,1bnq,1g1d,4qy3,1g52,2weh,1ttm,1avn,1lug,1ze8,1bnn,1xq0,3bet,3ryx,1if9,1oq5,2hl4,2osm,1g53,1bcd,3rz7,2weg,3rz0,2x7u,2hd6,2qo8,1i91,1bn1,1am6,1okl,1if7,1g54,4bf6,1if4,1bnw,2qp6,3rz8,1cin,1bnu,1okn,4bf1,1bn3,3mnu,2abe,3rz1,2hnc,3ibi,1if5,3ibn,1i90,1i8z,2f14,1cil,3ryj,4mdl,1xpz,4mdm,4mdg,4itp,4m2u,4kv0,3r17,4m2r,4ilx,4iwz,3r16

3 apo 3pxr,1pw2,1hcl

3 holo

3pxy,3py0,4ez3,1h0v,1jsv,2vtp,1jvp,1fvt,4kd1,3le6,2vti,3lfs,1pxi,4lyn,1ykr,1pxn,2vts,1p2a,2bhe,3unk,2vtq,2c68,3tiz,2c6o,2vtj,1w0x,2xmy,2uzo,1ke7,1dm2,1di8,1ke8,2c6i,1pxp,2uzn,2b55,3lfq,2w05,2xnb,1hck,3lfn,2a4l,1y8y,1pxj,2vto,2btr,2vta,3uli,2vth,1ke5,3fz1,1wcc,1pxo,3ti1,2c69,2vtm,2bts,1ckp,1ke6,2c6k,3ns9,2exm,4bgh,2r64,2c6l,2vtt,2vu3,2b54,2vtr,3unj,2a0c,3tiy,1vyz,3wbl,2vtn,1aq1,2c6m,2b52,1pxl,3s2p,1r78,2fvd,2vtl,1ke9,1pye,2b53,2duv,2c5y,2v0d,2w1h,1y91

4 apo 2a7h,1c5v,1c1o,2g55,1jrs,1tpo,1s0q,1c2m,2agi,1jrt,1tld,1c2l,2ptn

4 holo

1y3y,1o2i,1gi6,1o3d,1o39,1ql7,1c2i,1c2e,1yp9,1y3x,1oyq,1xug,1c5s,1xuh,2g8t,1xuk,1k1n,1c1t,2fx6,1mtw,1qa0,1o3e,1az8,1y3w,1qb6,1o3n,1c2k,1o3j,2bza,1c5p,1k1m,1mts,1o30,2agg,1s0r,1c5q,1ghz,1c5u,1n6x,1o2v,1gi4,1k1p,1o3g,2tio,1c2d,1c1r,1c2h,1o36,1o3l,1bju,1xuj,1tx7,1mtu,1o32,1o2q,1k1j,1c1s,1qbo,1k1o,1mtv,1g36,1o2l,1tx8,1gi1,1o3h,1o35,1ppc,1o2s,1o3m,1o2x,1o38,1bjv,1c1p,1c2j,1c2f,1o3i,1y3v,2blv,1o2p,1k1i,1gj6,1o2y,1o31,1xui,1c5t,1eb2,1o2k,1gi0,1

tps,1n6y,1o34,1pph,2g5n,1o2u,1o2r,1o2h,1qb9,2blw,1o3b,1o2z,1k1l,1o2o,1c1q,1c2g,1o37,1o3f,1o2n,1o3o,1ce5,1o2j,1aq7,1o3k,2g5v,1j8a,1qb1,1o2w,1xuf,1qbn,1o33,1c1n

5 apo

3dh5,4rat,1jvt,2g4w,2e3w,4j5z,2rat,1kf5,6rat,1rbb,4ao1,1rnu,1kf2,3euz,7rat,1kf8,3ev0,2rns,4j67,3i6f,4j64,1rnv,1jvv,3eux,5rat,1rph,3i7w,3ev1,4j65,3rn3,1fs3,8rat,4j68,4j62,1xps,1aqp,1rnw,3euy,1rbw,3rat,1bel,2g8q,3i6j,1js0,1rat,1rny,5rsa,4j6a,1rnx,3i67,2w5m,1rha,1rta,4mxf,9rat,4j63,1rnq,1rbx,2g4x,4l55,4j60,1rnz,1a2w,1rno,1kf3,4j61,1rtb,2blz,1xpt,1kf7,3i6h,4ot4,1afu,2blp,4j66,1rhb,3ev2,1kf4,7rsa

5 holo

1o0h,1jvu,3d7b,3d6q,3jw1,2w5i,3d8y,3d6o,3lxo,1o0o,1rpf,3d6p,4g8v,1rnc,1rob,1w4q,1o0f,1rnd,1rcn,1rnm,2w5k,1u1b,1qhc,1o0m,1eos,1w4o,1rnn,2xog,1w4p,3ev3,1rpg,1wbu,4g8y,1jn4,1o0n,1afk,2xoi,1eow,1f0v,2w5l,1z6s,4g90,2w5g,3d8z

6 apo

3msf,3ms3,3fxs,3fb0,4ow3,3fbo,3n21,4d91,4tnl,3p7s,2tli,5tli,4tli,7tli,3p7u,3p7v,3t2j,1fjo,6tli,1fjq,1fju,1fj3,1l3f,3t2i,2a7g,3eim,3p7p,1tli,3p7t,3p7q,8tli,3p7r,3p7w,3tli,2g4z,1fjv

6 holo

4mxj,1y3g,1kkk,1kr6,4n4e,3nn7,3msn,4mwp,1kro,1kjp,1kl6,4mzn,1kjo,1kto,4oi5,1ks7,4n66,1kei,4n5p,4mtw,1lnc,2tmn,1gxw,1pe8,1lne,1hyt,1lnb,2tlx,1pe5,1lna,3zi6,4tmn,5tmn,1zdp,1fjt,4tln,1qf0,1z9g,1qf2,1pe7,3tmn,8tln,1lnf,3f28,1qf1,1tlp,3fcq,1tlx,5tln,1tmn,3f2p,1thl,1os0,6tmn,1lnd

7 apo

4eyp,1trz,4f0o,3inc,1os4,3ilg,4fka,2vk0,4f1f,4ex1,4eww,3w80,2vjz,4ewx,4fg3,3w7y,4f0n,1g7a,4ey9,4f1g,4eyn,4ewz,4f1d,1guj,3q6e,3tt8,4exx,3ir0,3exx,1g7b,4f51,4f4t,4eyd,4f1b,4ex0,4f1a,4f1c,4f8f,4ey1,4f4v,1mso,1os3,3w7z,3e7z,4ak0,3e7y,2c8q,2c8r,1b17,3rto,4ins,1b2b,3mth,1b2e,1b19,2tci,1b2f,9ins,2g4m,1b2d,1b2g,1b2a,1b18,3i40,3ins,1b2c,1zni,4a7e,1m5a,3i3z

7 holo 2omi,2oly,1ben,1znj,2omg,1tyl,1tym,2olz,2omh,4ajx,4akj,1uz9,4ajz,1mpj,1wav

8 apo 3cbr,4mrb,2qgb,3w3b,1e3f,2g4g,1bmz,3a4d,1f41,1tta,3d7p

8 holo

4hju,2flm,3imu,2g5u,3cn3,3gs7,3imr,3kgu,3esp,3tct,2fbr,2f7i,4pm1,3ims,3d2t,4fi7,2qge,3ipe,1z7j,2b15,3cn1,2b9a,1tt6,3cn2,4fi8,2qgd,3eso,3glz,2g9k,3gs4,3ipb,1tyr,4hjt,1bm7,4l1s,1u2

1,2qgc,2b77,1e4h,3cn4,4ky2,3imv,2gab,3cn0,3gs0,4fi6,3m1o,4l1t,2rox,4i85,1e5a,3imw,1tz8,2f8i,3imt,2roy,3esn,3b56,1y1d

9 apo 1swa,1swb,1swc,3ry1,2rtb,2rtc,1slf,2rta,2ize,2izd,2izc,2iza

9 holo

2g5l,4cpf,1swe,4cpe,1swd,1mk5,2rtm,2rti,2rtj,2rtr,2rtd,2rtk,1lcz,2rtg,1sle,2rte,2rto,2rtf,2rtp,1sld,1slg,2rtq,2rtn,2rtl,2rth,1vwq,2izl,1vwf,1vwg,1vwc,1sts,2izi,1vwn,2izf,1vwr,2izh,1vwo,1vwk,1vwd,1vwl,1vwa,2izg,1vwh,1vwb,2izk,1vwe,1vwi,2izj,1str,1vwp,1vwj,1vwm,1srj,1srf,1srg,1sri,1pts

12 apo 4lyi,4ior,2oss

12 holo

4ioo,4hbx,4nue,4o7c,4ioq,2yel,4bw4,4o72,4ogi,4j0s,4f3i,4wiv,3p5o,4mep,4j3i,4lzs,4o71,4cfl,4c67,4meo,4o74,4bw2,3u5j,3u5l,4o7a,4nud,4o7b,4pce,4j0r,4mr3,4o78,4ogj,4mr4,4meq,4o77,4e96,4o7f,4cfk,4nuc,4a9l,4o76,4o75,3zyu,4ps5,4hbv,4hby,4pci,3mxf,4bw1,4men,4bw3,3u5k

13 apo 1kvm,2bls,1ke4

13 holo

4old,4jxw,2r9w,2pu2,1xgj,1xgi,4kz3,4kz7,1ll9,4jxs,2r9x,4kz4,1my8,4okp,4kz5,3gvb,4kz9,3gr2,4jxv,4kz6,3gqz

14 apo

1rps,2hbe,1cls,1hga,1hab,3d7o,2hbf,1rq3,1j3y,1yhe,1ljw,1xz2,1mko,1qsi,1hgc,1fn3,1kd2,1hgb,1j41,1xxt,1qsh,1dke,1bij,2w6v,1hho,1uiw,2dn2,2hhb,1gzx,2hbc,1bbb,1sdl,1bz0,2hbd,1hbb,1thb,1sdk,1yh9,2dn3,1j40,1ird,1a3n,1j3z,2hhd,1b86

14 holo 2d60,1g9v,2d5z

15 apo

2bdc,2fof,1h9l,2bd9,2foa,1esa,2foc,2foe,1lvy,1qix,2fo9,1gvk,2fob,2h1u,2fod,1haz

15 holo 1elc,1eld,2bd4,1elb,1ele,1nes,2bdb,1bma,2est,1ela,1qr3

16 apo 2pb8,2gns,2q1p,1cl5,1fb2,4gld,4fga,3fg5,2pyc

16 holo

2qvd,1zwp,1oxl,1tg4,1skg,1jq9,2arm,1fv0,1q7a,1tg1,1zyx,1kpm,1tp2,1tdv,1tj9,1jq8,1sqz,1zr8,1th6,1y38,3h1x,1sxk,1tk4,1sv3

18 apo 3lzy,4ape

18 holo

1e81,3er3,1epm,1eed,3pmy,4er4,3pww,3pmu,3er5,2v00,3pll,4er2,4lp9,3prs,1ent,1e80,3pbd,1od1,5er1,1epq,3pcw,3pgi,5er2,1epp,2er7,1epo,1er8,3pi0,1epl,1e5o,3pld,2er6,4lhh,1e82,1epn,2er9,3pm4,3pb5,3pbz,4er1

19 apo

1c6i,1c6e,1c6d,1l90,1c6c,1c6g,1c6f,1c6k,1c6h,3dmv,1c6j,2b6t,2b74,2b73,2b75,2b70,2b6x,2b6z,2b6y,2b72,2b6w

19 holo

3dn4,3hh3,3dmz,1l83,3hh6,1nhb,3dn2,3dn0,185l,186l,3hh4,182l,3dmx,188l,181l,3dn3,184l,187l,3dn6,3hh5,3dn1,183l,2otz,2oty,2ray,2raz,2rb2,2rb0

20 apo 1pud,4pun,1p0d

20 holo

1enu,2pwu,4q4s,2qii,2z7k,3sm0,4q4p,1k4h,3c2y,1r5y,3rr4,3eou,4q4q,1q66,1n2v,1q63,3gc4,1k4g,1p0e,1p0b,4q4o,3s1g,1s38,1q65,3eos,1f3e,4puj,2bbf,1s39,3tll,4puk,3gc5,3ge7,4q4r,1q4w

21 apo

2v1i,1azi,2frk,2o5q,3lr7,2vlz,1dwt,2frf,1npf,2o5s,2vlx,2frj,2o5t,1wla,3lr9,2o5b,1dws,2v1j,2v1f,2v1k,1gjn,2v1g,2fri,1ymb,2v1h,2v1e,2vm0,2vly,1dwr

21 holo 2o5l,2o58,2o5m,2o5o

22 apo 3n2d,3s9q,4l66,4kwn,4kmk,4jtb,3mrw,4kl4

22 holo 3n1n,3v2k,3rl9,3sj6,3u6z

23 apo 4e5b,1wfc,1r39

23 holo

3rin,1w84,1wbo,2zb0,1zz2,2zb1,1wbw,1w82,1w7h,1wbs,4e6c,1wbv,2zaz,4e6a,1wbn,1w83,1zyj,1wbt,1kv1,3kf7,3hv6,3hv5,3hv7,3gcu,4l8m,3huc,1m7q,1yqj,3itz,3gfe,3zya,3fc1,3u8w,1ouy,1ouk

24 apo 3m3d,1w75,1qif,1qie,1qih,1qig,2vt6,2vt7,1ea5,1qid,2va9

24 holo

1odc,1eve,1h23,2cmf,1h22,1dx6,1e66,1zgc,1zgb,1w4l,3zv7,2ckm,1w76,1w6r,1qti,1vot,1gpk,2ack,2xi4,2vjc,1gpn

25 apo 1mxw,1my0,1mxy,1mxv,1fto,1my1,1mxx,1mxz

25 holo

4o3b,1m5c,1my3,1wvj,1mqg,1m5e,3bki,2aix,3bfu,1ftl,1mqi,1ftj,1m5b,1syh,1my4,4igt,1fw0,4isu,1mxu,1mm7,4g8m,1nnp,4o3a,1nnk,1my2,1ms7,3bft,1ftm,1n0t,4o3c,1mm6,3tza,1mqj

27 apo

1dq0,1dq5,1nls,2g4i,2a7a,1dq2,3nwk,2uu8,1gkb,1jbc,3enr,1enq,1enr,1dq1,2ctv,1dq6,1qny,2enr,1scr,1apn,1vln,1nxd,1scs,1con

27 holo 1jn2,1cvn,4pf5,1gic,1jw6,1ona,1i3h,1hqw,5cna,3d4k,1qdc

28 apo

3snb,2gt7,3vb6,1uk2,2bx3,2a5a,3vb7,3snd,2c3s,1uk3,2gz9,2h2z,3vb3,2z3e,1uj1,2duc

28 holo 2gz7,3v3m,2z94,2gz8,1uk4

29 apo 1sug,3i80,1oem,1oes,2cm2,2cm3

29 holo

2cni,1nl9,2cm7,1ph0,1xbo,2cnf,1onz,2vev,1ony,1nny,2veu,3eax,1qxk,1pyn,1wax,2cnh,2vex,3eb1,2bge,1pxh,2vew,3i7z,2bgd,2vey,1nz7,2cng,2cm8,1no6,2cne,2cmc,2cmb

34 apo 2oxu,4ijo

34 holo

4gql,4efs,4gr3,1y93,3tsk,2oxw,3lik,4gr0,3ljg,1rmz,3lil,3lir,3ts4,3f16,3nx7,3n2u,3ehy,3f19,3ehx,3n2v,3f18,3f15,3f17,3lk8,3f1a

35 apo 4nva,1kxn,4oq7,4nvf

35 holo 4jpl,4jm6,4jqm,4jmz,4jm9,4jqj,4jqn,4jpu,4jpt,4jm8,4jm5

37 apo 3t0h,1yes,1yer

37 holo

3ft5,2xdx,2xht,2xjx,2xk2,2xhr,2xds,2xdk,2xab,2xdl,3ft8,4jql,3t10,4l94,3t0z,4l8z,4l91,3r4m,1byq,1yet

38 apo 3djh,3ce4,4p0h,4gru,3l5v

38 holo 1ljt,3ijg,1ca7,2ooh,2ooz,3dji,3u18,3l5r,3l5p,3l5t,3l5u,3l5s

39 apo

3dw1,3dvs,3de0,3de1,1bjr,3ptl,3i37,3de6,1ptk,3i2y,1egq,3de5,3de4,2pkc,3de3,3de7,3dvr,2prk,3d9q,3de2,3i30,3prk,3dwe,3ddz,3dw3,3dvq,3i34,4dj5,1cnm

39 holo 1pj8,1oyo

42 apo 3kbs,3qys,3kbv,1mnz,4e3v,3kbw,3kcj,2gub,4a8n,4a8i,4a8r,4w4q,2glk,4a8l,3kbj

42 holo 3kbm,4qeh,3u3h,4duo,4qe5,4qee,4qe4,4qe1,3kbn

43 apo 4ovh,1mmi,2pol,4pnv,3q4j,4k3s

43 holo

4mjr,4n94,4n98,4n95,4mjq,3d1e,3d1g,4k3o,4k3l,4n9a,4k3k,3d1f,4k3m,4k3r,4k3q,4k3p,4n99,4mjp,4n96,4n97

44 apo 2oot,2c6p

44 holo

3d7f,3bi0,4oc3,4oc2,2pvv,2pvw,4oc1,3d7d,4oc5,3d7g,3sjf,3bi1,2or4,3d7h,2xei,3bhx,4oc4,4oc0,3iww,2c6c,2jbj

45 apo 3the,4gwc,3tf3,4gsm,3th7,2zav,2pha

45 holo

3skk,3gn0,1wva,3kv2,2pho,3mfw,3lp4,3gmz,4gsz,3sjt,2pll,3mfv,4gsv,3dj8,3thh,3lp7,2aeb,4gwd

46 apo

2ofm,1ikj,1ywc,1d3s,1x8p,1np4,1ywd,1ywa,1eqd,4hpb,1x8q,1u0x,1x8n,1ywb,4hpa,1d2u,1x8o,1koi,3mvf,1erx

46 holo 4hpd,4hpc

50 apo 2y2v,3dl7,3dl4,2jgf

50 holo 4arb,4ara,4b80,4b82,4b7z,4b85,4b83,1n5r,2gyu,1j07,2gyw,2gyv,2jf0,2whq,2wu4

52 apo 3h2e,1tws

52 holo

4d8z,4db7,4d8a,4daf,3h22,4nil,1tx2,1tww,3h2m,3h24,3tyc,3h2f,1tx0,3h21,4d9p,3h26,3tyd,4dai,4nhv,3h2n,4nl1,3h23,3h2a

53 apo 3q40,2hd4,3osz,3dyb,2pq2,2duj,3aj8,3q5g,3qmp,1ic6,4b5l,3aj9,2v8b,2g4v,2dqk

53 holo 1p7v,1p7w

55 apo 2h7q,1phc

55 holo

1re9,1noo,4cp4,2cpp,6cpp,1yrc,8cpp,3cpp,3cp4,5cpp,5cp4,1phe,7cpp,1yrd,4cpp,2z97,1iwi,2zax,2zwu,2zaw,2zwt

56 apo 20gs,16gs

56 holo

1aqx,3csj,10gs,2gss,7gss,1aqw,3gss,9gss,1aqv,5gss,6gss,8gss,2a2s,12gs,11gs,3pgt,2a2r,18gs,3dgq,3dd3

57 apo 3wc6,1z5r

57 holo

2piz,1zg7,2pj4,2pj8,2pj9,1zg9,2jew,2pj5,3wc5,2pjc,2pja,1zg8,3wc7,2pj2,2pj3,2piy,2pj0,3wab,2pjb,2pj1,2pj6,2pj7

58 apo 2zk0,3vsp

58 holo 3an3,2zno,3vjh,3d6d,3vso,2i4j,2zk6,3an4,2i4p,3hod,2i4z,3r8i,4e4q,2yfe

60 apo 1uuw,1o7w,1o7m,1ndo,1o7h

60 holo 1o7g,2hmm,1uuv,2hmo,1eg9,1o7n,2hmk,1o7p

61 apo 3o5p,3o5q,3o5m,3o5l,3o5o

61 holo

4jfk,4jfj,4dro,4drp,4tx0,4drq,4w9o,4tw6,4jfl,4w9q,3o5r,4w9p,4jfm,4drm,4drk,4drn,4tw7

63 apo 3p6r,3l61,3l62,3p6w,3p6u,3p6v,3p6s,3p6q

63 holo 1t87,4g3r,1t88,3l63

64 apo 3o7s,1ier,3u90,3af7,3f32,2z5r,2zg7,1gwg,2w0o,2z5q

64 holo 3f39,1xz1,3f35,3f33,3f38,3f37,3f34,3f36,1xz3

65 apo 4q6h,4e35,4e34

65 holo

4nmt,4k6y,4k76,4joe,4joj,4jor,4jog,4k72,4nmo,4jok,4jop,4joh,4nmp,4nmq,4nmr,4nmv,4jof,4nms

66 apo 1bsq,4lzv,3npo,4gny,4lzu,4iba,4ib9,1b8e

66 holo 3nq3,1gx8,3uex,1b0o,3uev,1gx9,3uew,3nq9,3ueu,1gxa

67 apo 1mop,2a88

67 holo 2a84,1n2o,3le8,3isj,1n2g,1n2h,4ef6,1n2i,2a7x

68 apo 4jtn,2b5h,4iep,4iez,4ieq,4kwj,4iex,4ieo,2gh2

68 holo 4kwk,4iev,4ies,4iet,3eln,4jto,4ier,4kwl,4iew,4iey,4ieu

69 apo 1d7h,1d6o,2ppn

69 holo 1d7j,2fke,1qpf,1fkj,1j4h,1fkf,2dg3,1fki,1fkh,1a7x,1d7i,1fkb,1fkg,1j4i,1fkd,1j4r

70 apo 1cmt,1cmq,1aa4

70 holo
2eus,2euu,1cmp,2as6,2eup,1ryc,2as3,2eun,2as2,2euq,2as1,2y5a,2eut,2euo,2aqd,2eur
72 apo 1aee,1aes
72 holo
1aeh,1aeu,1aed,1aek,1aev,1aeb,1aet,1aej,1aeo,1ac8,1aeq,2anz,1aef,1aem,1aen,1aeg,1ac4
76 apo 2tga,1tgt,1tgb,1tgn,1tgc,2tgt,1btp
76 holo 1tnh,1tnl,1tnk,1tng,1bty,1tni,1tnj
78 apo 3fvj,4p2p
78 holo 2azy,2b01,2azz,2b04,2b00,3o4m,2b03,1l8s,1fxf,1fx9,3qlm
79 apo 3dr9,2qfk,4dwt,4dwu,3ord,3kun,4jyq,3mou,1ew6,4gzg
79 holo 3lb3,3lb1,4fh7,4fh6,3lb2,4ilz,3lb4
80 apo 8rnt,9rnt,3rnt
80 holo 6gsp,5gsp,6rnt,1rnt,3bu4,1i0v,2rnt,2bu4,5bu4,1rgc,1gsp,7gsp,1bu4,4bu4
81 apo 4qut,4tt2,4qsr,4tu6,3dai,4qsq
81 holo 4tu4,4tz8,4tte,4tyl
82 apo 4fov,2vw2,2vw0
82 holo 4fpe,4fph,4fpf,4foq,4fpc,4foy,4fpl,4fow,4fpj,4fpk,4fq4,4fpy,4fp2,2vw1
85 apo 1w5z,1w6c,1w6g,1sii,3kn4,1sih,1rjo,1w4n,3kii
85 holo 2bt3,2cg0,2cg1,2cfw,2cfg,2cfl,2cfd
86 apo 1qhz,7a3h
86 holo 1qi0,1e5j,1hf6,8a3h,1w3l,4a3h,1h2j,1ocq,1w3k
88 apo 3c2x,3umq,3usx,3uil,3uml,4q9e,3qs0,4oug,4q8s,4orv,4fnn
88 holo 3rt4,3nw3,3o4k
91 apo 4g03,2vue,1ao6,4emx,3jry,4g04,1bm0
91 holo 2xw1,4l8u,4l9k,4iw2,4la0
92 apo 1hai,1abj,1ppb
92 holo 1hxf,1qhr,1qj1,1c5l,1hxe,1qj6,1awf,1ny2,1qj7,1ad8,2uuf,1hah
93 apo 3sgx,1ueh,3qas
93 holo 3th8,1x07
94 apo 3p1e,3dwy
94 holo 3p1d,4a9k

97	apo	1arp,1gzb,2e39,1arw,1arx,1ary,1gza,1arv,2e3a,1aru
97	holo	1hsr,1ck6
98	apo	1gdu,1xvm,1gdq,2g51,2g52,1xvo
98	holo	1gdn,1fy5,1fy4,1pq8,1fn8,1pq7,1pq5
99	apo	3ux0,3p1n
99	holo	4dhs,4dhp,4dhu,4dhq,4dhr,4dho,4dhn,4dht,4dhm
100	apo	2dea,3fh4,2nyq,1rtq,1lok,2prq,1amp
100	holo	1txr,3vh9,1igb,1cp6,1ft7,2iq6
101	apo	4bro,4brp,4brm,4br9
101	holo	4bra,4bri,4brl,4brq,4brf,4brh,4brn,4brc,4bre
102	apo	1fz0,1fz2,1fz4,1fz7,1fyz,1fz6,1fz9,1fz5,1fz8,1fz1,1fz3
102	holo	1xu3,1xvd,1xu5
104	apo	1xll,1xlh,1xlb,1xla,1xlk,1xle
104	holo	1did,1xlc,1xlj,1die,1xlf,1xli,1xlg,1xld
105	apo	1gmd,1ab9,1gmc,2gch,1yph,2gct,1gct
105	holo	2p8o,1gha
106	apo	1bs5,1xen,1bs4,1icj,1xeo,1bsz,1xem,1bs7
106	holo	1g2a,1bs8,2ai8,1g27,1lru,1bs6
108	apo	1q9k,1q9l
108	holo	2r2e,3sy0,2r1y,2r2b,2r1x,3t65,2r1w,3t4y,2r2h,3bpc
109	apo	2iuv,2iur,2ah1,2iup
109	holo	2agw,2iuq,2hjb
110	apo	4jy0,3mwv
110	holo	4ju6,4j06,4j08,4j0a,4jvq,4jjs,4ju7,4ju3,4jju,4ju4,4j02
111	apo	1dv1,1bnc
111	holo	2w6o,2v59,2w70,2w6p,2v5a,2w6n,2w6m,2w6q,2w6z,2v58,2w71
112	apo	3yas,1qj4,2yas,6yas,3c6y,3c6x,7yas,3c70,3c6z
112	holo	1yas,1sc9,5yas
113	apo	4afn,4bnw
113	holo	4bnt,4bny,4bnv,4bo1,4bo0,4ag3,4bo2,4bnx,4bnu,4bo3,4bnz
116	apo	3dps,1y1t,1sj9,2hsw,1y1q,3ddo,1y1r

116	holo	4e1v,3fwp
118	apo	1h05,2dhq
118	holo	2y76,4b6o,2y71,4ciw,4b6p,4ciy,1h0r,2xb8,1h0s,4b6q,2y77
121	apo	2whe,1zol
121	holo	4c4r,1z4o,4c4s,4c4t,1o03,1z4n
122	apo	4v2f,1bjz
122	holo	1ork,2x9d,2fj1,2vke,2o7o,2x6o
123	apo	1td1,3e9z,1tcu
123	holo	3fb1,3e9r,3e0q,3fnq,3faz,3iex,3djf,3f8w
124	apo	1tk3,1nu6
124	holo	3kwf,4lko,4jh0,2ogz,2ole,1rwq,3eio,1n1m
126	apo	2e3m,2e3s
126	holo	3h3r,2e3r,3h3s,2e3p,2z9y,3h3q,2e3n,2e3o,3h3t,2e3q
128	apo	1fx6,1fxp
128	holo	1fy6,3e12,1fwn,1pck,3e0i,1pe1,2a21,1jcx,1fws,1pcw
130	apo	2cz7,2cyz,2zpb,2ahj
130	holo	2zph,2zpf,2zpe,2zpi,2zpg
131	apo	3tpj,3tpl
131	holo	4ivt,4dv9,3tpp,4fgx,4dvf
132	apo	3gxi,2nt0,1ogs,3gxd,2f61,3gxm,2nt1
132	holo	3gxf,2nsx,3rik,3ril
137	apo	3bui,3bud
137	holo	3cv5,3bvw,3czn,3czs,3bvv,3buq,3bvx,3bvt,3bvu
138	apo	8adh,1ye3
138	holo	2jhg,2jhf,1n92,1hld,1het,1heu,2ohx,2oxi,1hf3
139	apo	4usv,4clf,4cll,4cls,4clu,4ust
139	holo	4clw,4clk,4usu,4usw
141	apo	1hcb,1hug,1huh
141	holo	2nmx,1bzm,2fw4,3lxe,2nn7,1czm,2nn1,1azm
142	apo	1ous,1oux
142	holo	1uzv,2bp6,3dcq,1our,2boj,1ovp,1ovs

143	apo	1swx,3rwv
143	holo	4gjq,2evl,3s0k,4gix,4gxg,3rzn,2euk,4h2z,1sx6
144	apo	3iqo,4pdz,3cr2
144	holo	3gk2,3gk4,3gk1
146	apo	3fwq,3at2
146	holo	3bqc,3c13,3at4,3axw,2zjw
148	apo	1ozw,1s13,1twn,1ozr,1s8c,1t5p,1n45,1twr
148	holo	3hok,3k4f
149	apo	1ejd,3spb,1ejc
149	holo	3upk,1eyn,3lth,3swq
150	apo	1m35,2bhb,1wl6,2bhc,1wlr
150	holo	2bhd,1a16,1n51,2bh3,2bha,2bn7
152	apo	2fu6,2fm6,2h6a,1sml,2fu7
152	holo	2gfj,2gfk,2aio,2qdt,2fu8,2fu9
153	apo	2q2m,2q39,4ib8,2blg,1qg5,1bsy,4ib7
153	holo	4ib6,1bso
154	apo	2vb9,2buh,1g5x
154	holo	1fj4,2aq7,2vb8,2aqb,2vba
157	apo	3vaj,3vam,3val,3vai,3vag,3vaf,3vak,3vah
157	holo	4tu9,4tu8
158	apo	3cdn,3cab,3cz2
158	holo	3bfb,3bfh,3bfa,3bjh
159	apo	2b3l,2b3h,2b3k
159	holo	2nq6,4ikt,4ikr,4iku,2nq7,4iks
161	apo	1bbc,3u8b,1pod
161	holo	1poe,1db4,1kqu,1kvo,1j1a
162	apo	2r2x,1rtc
162	holo	4mx5,1br6,4hv3,4mx1,2p8n,3hio,2pjo,2r3d
163	apo	2y1e,2jd2,2y1c
163	holo	3zhz,3zi0,2jd1,2y1d,2y1g
167	apo	1znk,2ozq,1qy0

167	holo	1znh,1zne,1znl,2dm5,1qy1,1zng,1qy2
168	apo	3qcb,3qcd,3qcc
168	holo	3qch,3qci,3qcj,3qce,3qcf,3qcg,3qck
169	apo	1bhs,3km0
169	holo	1jtv,1dht,3hb4,1i5r,1qyv
170	apo	1meo,4ew1
170	holo	1rbm,1rbq,4ew3,4ew2,1rc1,1rc0,1njs,1rbz
171	apo	1r3m,1r5d,1bsr
171	holo	11ba,3djo,3djv,1r5c,3djq,3djp,3djx
172	apo	3f0d,3f0f,3f0e
172	holo	3k2x,3f0g,3ieq,3jvh,3k14
174	apo	3ojn,3ojj,3ojt,2ig9,3bza,1q0o
174	holo	4ghh,3ojk,1q0c,4ghg
175	apo	1w8v,2cpl,3k0n,3k0m
175	holo	1w8l,1vbt,1ynd,1nmk,1vbs,1w8m
178	apo	1ivg,1nn2
178	holo	1ing,1inh,1ivf,1ive,1inx,1inw,1ivc,1ivd
179	apo	1gqv,1hi2
179	holo	2c05,1hi3,2c01,2c02,1hi5,2bzz,1hi4
180	apo	4i2g,4i2a,4i2f,4i29
180	holo	4i2d,4i2i
181	apo	3g46,1nxf
181	holo	3g4w,3g4v,3g4u,3g52,3g4y,3g53,3g4r
182	apo	1tpd,2v5l,1ag1,1tpf,5tim
182	holo	1iih,4tim,6tim,1trd
183	apo	2bno,1zz6,1zzc,1zz9,2bnm
183	holo	1zz8,1zz7,2bnn,1zzb
184	apo	2cxn,2cvp,2cxu
184	holo	2cxs,2cxr,2cxq,2cxo,2cxt,2cxp
186	apo	1jp7,1jdu,1jds,1jpv,1je0
186	holo	1jdv,1jdt,1je1,1jdz

188	apo	2ij5,1n40,3g5f
188	holo	4g46,1n4g,4g48,4g47,2ij7
189	apo	2x16,2x1u
189	holo	2x1t,2vel,2ven,2x1s,2x1r,2x2g
190	apo	1lls,1jw4,1omp
190	holo	1jw5,1anf,3mbp,1dmb,4mbp,1ez9
191	apo	3zty,3ztw,3zwk
191	holo	3zup,3zwd,3zx5,3zw7,3zx4,3zu6
192	apo	4d98,4d8y
192	holo	4dao,4da6,4da7,4dae,4d9h,4d8v,4dab
196	apo	1l4b,1l5n
196	holo	1jhm,1jhu,1jh8,1jhp,1jhx,1l4e,1l5f
197	apo	2hcv,3iud,3iui,3itx,3iuh,3ity
197	holo	2i56,2i57
199	apo	1o7j,1hfj,1hfk
199	holo	1hg0,1hfw,1hg1
202	apo	2v7i,2x66
202	holo	2x67,2v7k,2v7m,2v7l,2x68
203	apo	1w6l,1w8e,1gsk,2x88,1w6w,2bhf
203	holo	1of0,3zdw
205	apo	1sjs,3icd,1pb3
205	holo	1pb1,5icd,1p8f,9icd
207	apo	3q6q,3q6r
207	holo	3q3a,3pa2,3q38,3ry8,3pa1,3q39
208	apo	1pop,1ppd
208	holo	1pip,1bqi
209	apo	1dea,1cd5,1fs6,1fsf
209	holo	1hor,1frz,1hot
211	apo	3i13,1bvt,3i11,1bc2
211	holo	1mqo,4tyt
212	apo	3l3i,3l3g,3l3d,3l3j

212	holo	1m6o,3kpm,3dx6,3kpl
216	apo	1dhy,1eil,1kw3,1eiq
216	holo	1kw6,1kw8,1eir,1kw9
217	apo	1tew,1jse,2lz2,135l,3lz2
217	holo	1lzy,1jef,1ljn
220	apo	1zah,1zal,2quv
220	holo	2ot1,1zaj
221	apo	4lgz,3v9j
221	holo	3v9l,4lh1,3v9k,4lh2,4lh3
222	apo	1iae,1iaa,1iad,1ast,1iab,1iac
222	holo	1qjj,1qji
224	apo	1mr5,1ms3,1ms4
224	holo	1ms8,1ms9
228	apo	1dup,1eu5,1euw
228	holo	1seh,1dud,1rn8,2hr6,2hrm
229	apo	4jh2,4jh1
229	holo	4jh4,4jh7,4jh3,4jh6,4jh9,4jh5
232	apo	2i3u,2i3r,2i4e
232	holo	2i5x,2i4h,2h02,2i4g,2h04
235	apo	2fs6,2fs7
235	holo	2cbs,1cbs,2fr3,3cbs,1cbq
236	apo	3d5g,1py3,1pyl
236	holo	3dgy,3d4a,3d5i,3dh2
237	apo	1ey0,1eyd,1stn
237	holo	4wor,1sth,1snc,1stg
239	apo	4m5v,4m5q
239	holo	4m4q,4mk5,4m5u,4w9s,4mk2
241	apo	3f1n,3f1p
241	holo	3h82,4gs9,3h7w,4ghi,3f1o
244	apo	4lsf,4lse,4lsi,4lsh
244	holo	4gcs,4gcq,4gcp

245	apo	3b4o,3ex9
245	holo	3cnm,3jum,3juo,3b4p,3jup
246	apo	1dcs,1rxf
246	holo	1uog,1uof,1rxg,1uo9
247	apo	3gte,3gke
247	holo	3gl2,3gob,3gl0,3gb4,3gts
248	apo	1rrh,1rrl
248	holo	1n8q,1jnq,1hu9,1no3
249	apo	3hcv,3czf
249	holo	1w0w,3bp7,1uxw,3b3i,1of2
252	apo	1gpf,1e15
252	holo	1w1v,1e6r,1o6i,1w1p,1ur8
253	apo	1rtm,1kwt
253	holo	1kww,1kwy,1kwv,1kwx,1kwu
255	apo	1l7l,1uoj
255	holo	4lkf,4lk7,4lkd,4lke
257	apo	1ofb,1ofp
257	holo	1oab,1ofo,1of6,1ofa,1hfb
258	apo	3wng,3wne,3wnf,3wnh
258	holo	3vqe,3vq8,3vq5
259	apo	1f7d,1f7o
259	holo	1f7k,1f7r,1f7n,1f7q,1f7p
260	apo	4fot,4hoy,4iko,4fop,4jy7
260	holo	4jwk,4jx9
261	apo	1une,2bpp,1bp2,1g4i,1mkt
261	holo	1fdk,1mkv
262	apo	4erx,4qaj,4fno,4jc4
262	holo	4qd3,4qbk
263	apo	2q08,2q6e
263	holo	3hk9,3hk8,3hk7,3hka,3hk5
264	apo	1alb,1lib

264	holo	1lid,1lif,3jsq,3hk1,2ans
266	apo	1t8p,3nfy
266	holo	2f90,2a9j,2h4x,2h4z,2h52
269	apo	1n9k,3cz4,1n8n,1rmq
269	holo	2g1a,1rmt,1rmy
270	apo	1pvf,2vnq
270	holo	2vnp,1nfs,1ppw
271	apo	4rj2,3onv,3ooe
271	holo	1pke,1pk9,1pk7,3ut6
273	apo	3wra,3wr8
273	holo	3wr3,3wr9,3wr4,3wpm,3wrc
276	apo	1yze,1yy6
276	holo	2fop,2foo,2foj
279	apo	4e52,1pw9
279	holo	3ikr,3ikq,3ikp,3ikn
280	apo	3rdu,3rds,3rdx
280	holo	3rdq,3rdo,3rdm
282	apo	1m4s,1m4t
282	holo	2vu0,1nl7,1ou6,2vu1
286	apo	2yeu,2yf4,2yf9,2yf3
286	holo	2yfd,2yfc
287	apo	4bt7,4bt2,4bt6
287	holo	4bt3,4bt5,4bt4
288	apo	2afm,2afo
288	holo	3pbb,2afw,2afz,2afx
291	apo	4i64,1r7i
291	holo	4i6a,4i6w,4i56,4i6y
292	apo	1gmq,1sar
292	holo	2sar,1rsn,1gmr,1gmp
293	apo	2fmi,2fmf,2flk
293	holo	2fmh,2fka,2flw

294	apo	1n1b,1n1z
294	holo	1n22,1n20,1n23,1n24
296	apo	3kqz,3kqx
296	holo	3t8w,4r76,3kr4,4k3n
297	apo	1hn9,1hnk
297	holo	1hnd,1hnj
298	apo	1f6k,1f6p,1f5z
298	holo	1f73,1f74
300	apo	1y6v,1ed9,3tg0,1ed8
300	holo	1ew9,1ew8
301	apo	1mkx,1mkw
301	holo	1uvt,1ett,1etr,1ets
303	apo	2aju,2ajs
303	holo	2ajx,2ajz,2ak1,2ajv
305	apo	3s0f,3s0a,3rzs
305	holo	3s0e,3s0b,3s0d
306	apo	3par,3pbf,1r14,1r13
306	holo	3pak,3paq
307	apo	4enl,3enl
307	holo	5enl,1els,6enl,7enl
310	apo	3ppg,3ppf
310	holo	4l61,4l64,4l5z,4l6o
313	apo	1eo2,2bum,1eoa
313	holo	2bur,1eob,1eoc
315	apo	2ptx,2ptw
315	holo	2ptz,2pu1,2pty,2pu0
318	apo	1v8n,1v8i
318	holo	1v8s,1v8l,1v8m,1v8r
320	apo	2de6,2de5,3vmh
320	holo	2de7,3vmg,3vmi
322	apo	1u1s,4j6y,1u1t

322	holo	4j6x,4j5y
326	apo	3kdh,3kl1
326	holo	3kdi,3nr4,3ns2
327	apo	3k5t,3hi7,3mph
327	holo	3hig,3hii
328	apo	2psr,1psr,3psr
328	holo	2wos,2wor
329	apo	1ime,1imf,4as4
329	holo	1ima,1imb
331	apo	3m8y,3m8w,3twz
331	holo	3ot9,3uo0
332	apo	1ho1,1ixq,1ixp
332	holo	1ho4,1ixo
334	apo	2gqv,1vie,2rh2
334	holo	2rk2,1vif
339	apo	1arl,1m4l,1cpx
339	holo	1f57,2rfh
342	apo	1yfu,4hvo
342	holo	1yfx,1yfw
343	apo	3ssw,1mi7
343	holo	1wrp,2oz9,1zt9
345	apo	3zo8,1dbf,2chs
345	holo	1com,2cht
346	apo	1v3b,1v2i
346	holo	1v3c,1v3d,1v3e
347	apo	4j0d,4j0c
347	holo	4j0h,4j0i,4j0j
348	apo	2isd,1djh,1dji
348	holo	1djx,1djw
349	apo	2oa6,2e4o
349	holo	3bnx,3cke,3bny

351	apo	1fxx,3c95
351	holo	3hl8,2qxf,3hp9
352	apo	2sga,1sgc,3sga
352	holo	5sga,4sga
354	apo	4j0u,3zgq
354	holo	4hor,4hot,4hos
355	apo	4h00,4h01
355	holo	4lcq,4lcr,4lcs
359	apo	2oqy,3gd6,3fyy
359	holo	3es8,3es7
360	apo	3wsi,3wse,3wsd
360	holo	3wsf,3wsg
362	apo	4itn,4ehx
362	holo	4itm,4ehy,4itl
363	apo	3wmy,3wmz
363	holo	3wn2,3wn1,3wn0
364	apo	1ttc,3djt,3kgs
364	holo	3kgt,1eta
366	apo	3a7g,3a7f
366	holo	3a7h,3a7j,3a7i
367	apo	2y40,2y3z
367	holo	2y42,2y41
368	apo	3azy,3b01,3azx
368	holo	3b00,3azz
372	apo	1qdt,1qus,1qdr
372	holo	1d0l,1qut
375	apo	3pn2,3m6o
375	holo	3m6p,3pn4,3pn3
378	apo	1byi,1dbs
378	holo	1dai,1dae,1dad
380	apo	1nx2,1alv,4phn

380	holo	1nx3,1alw
381	apo	4a8g,4a88
381	holo	4a87,4a85,4a80
384	apo	2wlc,2wld
384	holo	2wle,2wlg,2wlf
387	apo	1paf,1qcg
387	holo	1d6a,1qcj,1qci
388	apo	3pb6,3pb4
388	holo	3pb7,3pb8,3pb9
389	apo	1lbv,1lbw
389	holo	1lby,1lbx,1lbz
390	apo	2pa5,4icz
390	holo	4ge6,4ge5,4ge2
391	apo	1u6j,1u6i
391	holo	3iqz,3iqf
392	apo	2dup,2duo
392	holo	2dur,2duq
393	apo	3n5k,1su4
393	holo	2agv,1wpg
395	apo	3per,3q1g
395	holo	3pf7,3pm5
396	apo	1gy0,1gxy
396	holo	1gxz,1og1
398	apo	1ecg,1ecf
398	holo	1ecj,1ecc
401	apo	2d6k,2d6l
401	holo	2d6n,2d6m
402	apo	4bwo,4b4p
402	holo	4b4q,4b4r
405	apo	2ys7,2yrw
405	holo	2ys6,2yrx

408	apo	2o9z,2oam
408	holo	2e4g,2oal
409	apo	3mwt,3mwp
409	holo	3mx2,3mx5
410	apo	1dc5,1dc3
410	holo	1gad,1dc6
414	apo	1ogf,1ogc
414	holo	1oge,1ogd
415	apo	1oxt,1oxs
415	holo	1oxu,1oxv
418	apo	3mvi,3mvt
418	holo	1add,1a4m
419	apo	1j1q,1gik
419	holo	1j1r,1j1s
420	apo	3ivy,3iw0
420	holo	3iw2,3iw1
421	apo	2ris,1tks
421	holo	2riu,1tku
422	apo	3o4g,1ve6
422	holo	2hu5,2hu7
423	apo	1m47,1m4c
423	holo	1m48,1m49
424	apo	2yvm,2yvn
424	holo	2yvp,2yvo
425	apo	3e8m,3e84
425	holo	3e81,4hgo
428	apo	1sz8,1yxl
428	holo	1td7,1oxr
430	apo	3zdx,3t3p
430	holo	3zdy,3ze2
431	apo	2j46,2j45

431	holo	1o87,2c04
433	apo	1j2t,1j2u
433	holo	1v7z,3a6d
435	apo	2z3g,1wn5
435	holo	2z3h,1wn6
437	apo	7xim,4xim
437	holo	1xim,6xim
438	apo	4by3,4c6c
438	holo	4c6m,4c6l
439	apo	2y7e,2y7d
439	holo	2y7f,2y7g
440	apo	4w5h,4ntz
440	holo	4w5j,4nu0
442	apo	1ogh,1pkh
442	holo	1pkk,1pkj
444	apo	1odl,1odk
444	holo	1odi,1odj
445	apo	4kjt,4hsw
445	holo	4kmv,4hsx
447	apo	1u98,1u94
447	holo	1xms,1xmv
449	apo	3dxl,3dy9
449	holo	3dye,3dzt
451	apo	3kx7,3kv7
451	holo	3kvz,3kw1
452	apo	1ah6,1ah8
452	holo	4asa,4asg
453	apo	2ou4,2qul
453	holo	2qun,2qum
455	apo	3lxh,4c9m
455	holo	3lxi,4c9l

457	apo	1x8g,3f9o
457	holo	2gkl,2qds
459	apo	3ppp,3ppn
459	holo	3ppr,3ppq
460	apo	2f2q,1t8a
460	holo	2f47,2f32
461	apo	1is6,1is5
461	holo	1is3,1is4
462	apo	3kje,3kjh
462	holo	3kjg,3kji
463	apo	1q7z,1q7m
463	holo	3bol,3bof
465	apo	3t2f,3t2b
465	holo	3t2d,3t2e
466	apo	1gsd,1k3o
466	holo	1k3l,1k3y
467	apo	2hlh,2hhc
467	holo	3six,3siw
468	apo	1knq,1ko1
468	holo	1ko5,1ko8
471	apo	1yuy,1yv2
471	holo	1yvx,1yvz
472	apo	2g3x,2noy
472	holo	4i89,4i87
473	apo	3quq,3qx7
473	holo	3qu2,3qxg
475	apo	1adi,1ade
475	holo	1hon,1hop
476	apo	3jsl,3jsn
476	holo	4cc6,4cc5
478	apo	4r05,4r8r

478	holo	4r8s,3p97
480	apo	1lci,1ba3
480	holo	3rix,4e5d
484	apo	3ca0,3c9z
484	holo	3ca6,3cah
485	apo	1e43,1e3x
485	holo	1e3z,1e40
489	apo	1v3r,1vfj
489	holo	1v3s,1v9o
491	apo	1d9e,1x8f
491	holo	1phw,1g7v
493	apo	1k6i,1k6j
493	holo	1k6x,1ti7

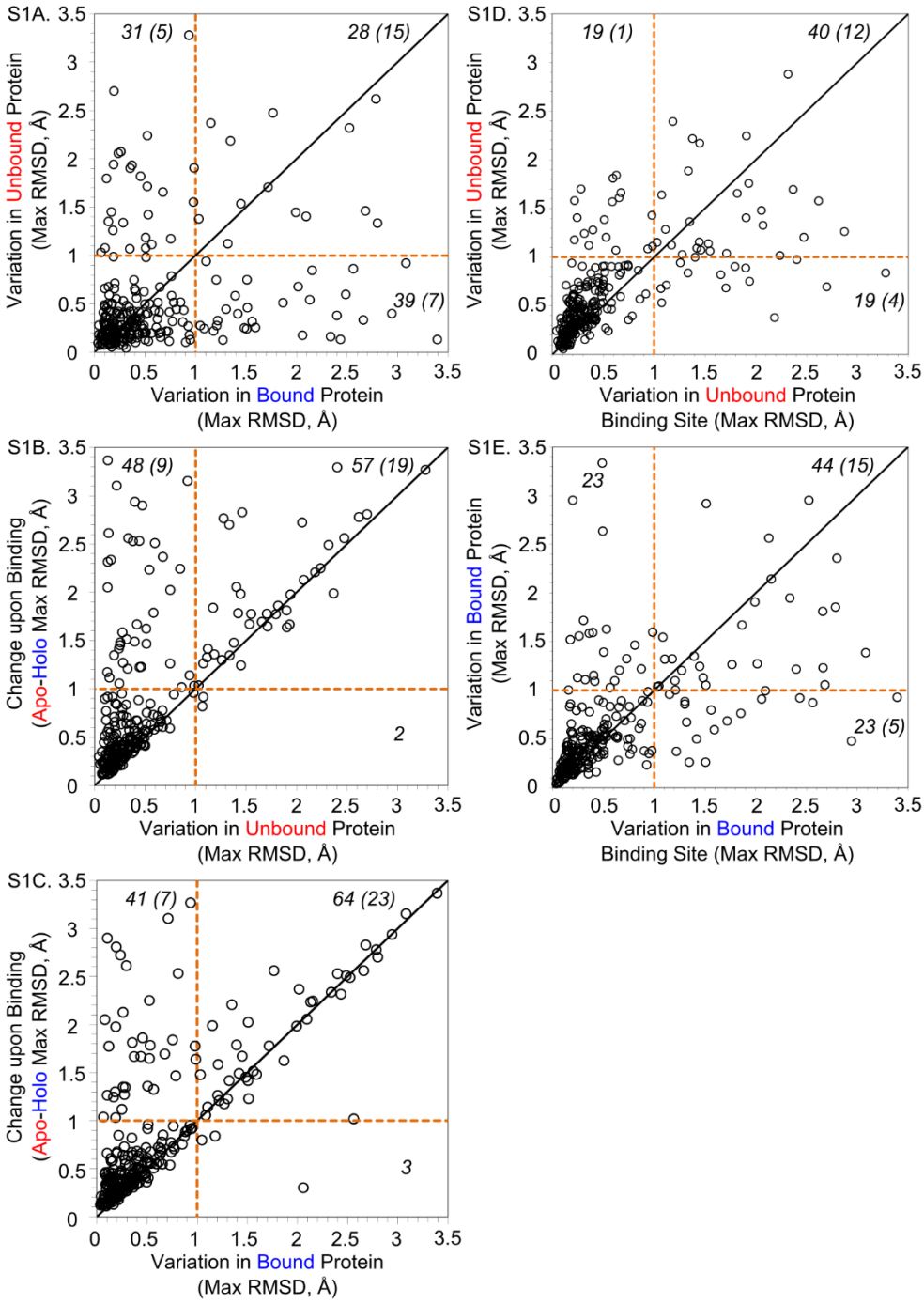
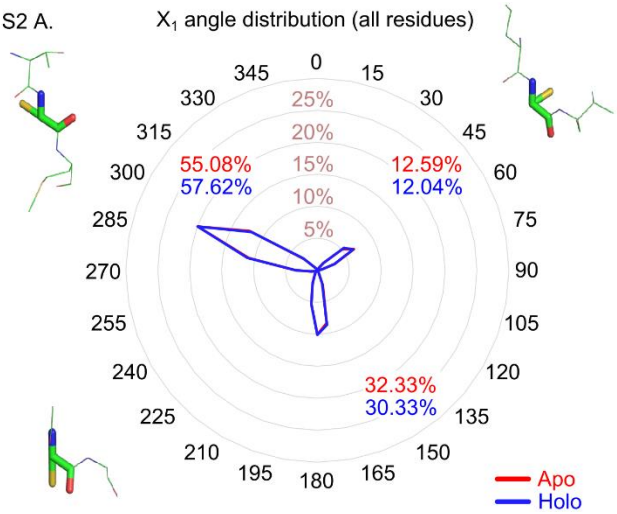


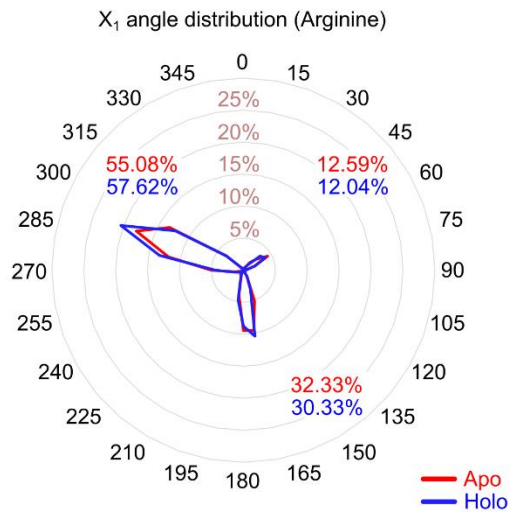
Figure A-1 (S1A-S1E). Analyses of maximum backbone RMSD for only unified binding site residues within each protein family. Each point represents the maxima observed in one protein family, and the number of points of each section is labeled in black (numbers in parenthesis are points with values > 3.5 Å). A) The maximum across the apo-apo pairs is compared to the maximum of the holo-holo pairs, binding site residues only; 207 proteins display RMSD ≤ 1 Å for both groups. B) The maximum across the apo-holo pairs is compared to the maximum of the apo-apo pairs, binding site residues only; 201 proteins display RMSD ≤ 1 Å for both groups. C) The maximum across the apo-holo pairs is compared to the maximum of the holo-holo pairs, binding site residues only; 201 proteins display RMSD ≤ 1 Å for both groups. D) The maximum across the apo-apo pairs for only binding-site residues is compared to the whole backbone maximum for apo-apo pairs; 227 proteins display RMSD ≤ 1 Å for both groups. E) The maximum across

the holo-holo pairs for only binding-site residues is compared to the whole backbone maximum for holo-holo pairs; 214 proteins display RMSD ≤ 1 Å for both groups.

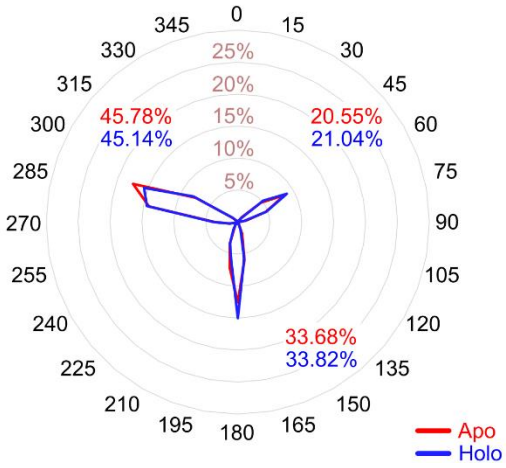
S2 A.



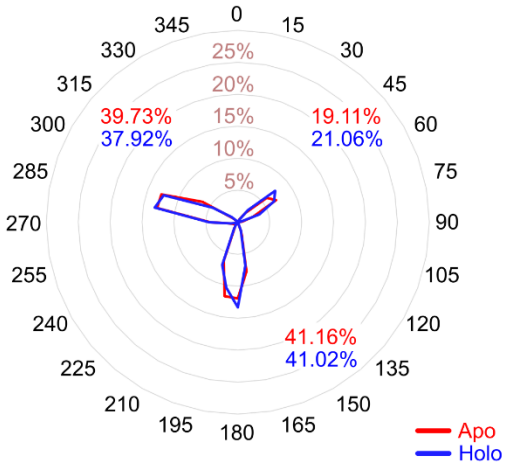
S2 B.



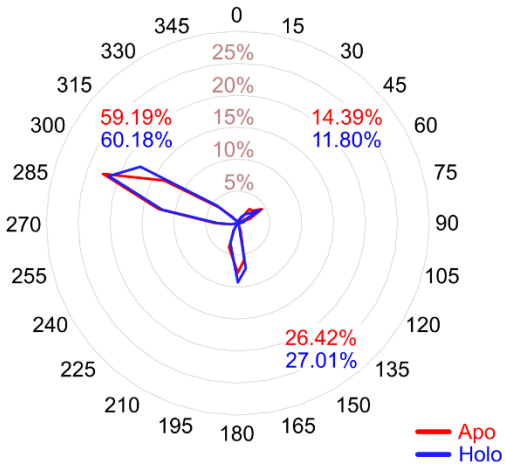
S2 C.



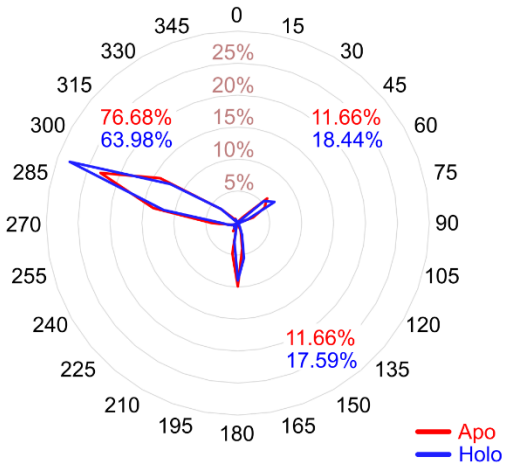
S2 D.



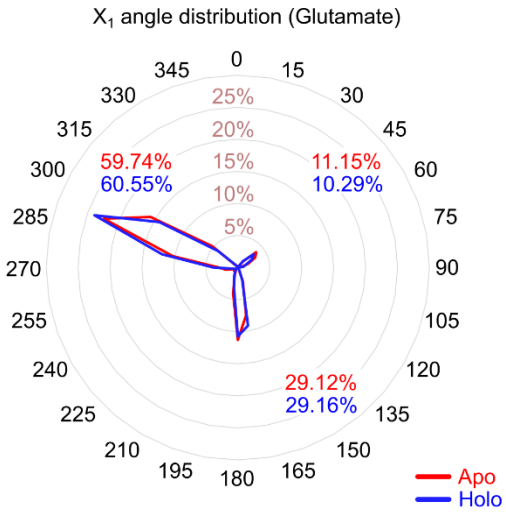
S2 E.



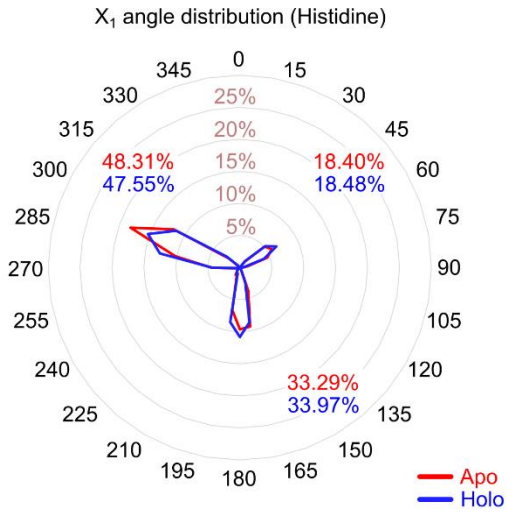
S2 F.



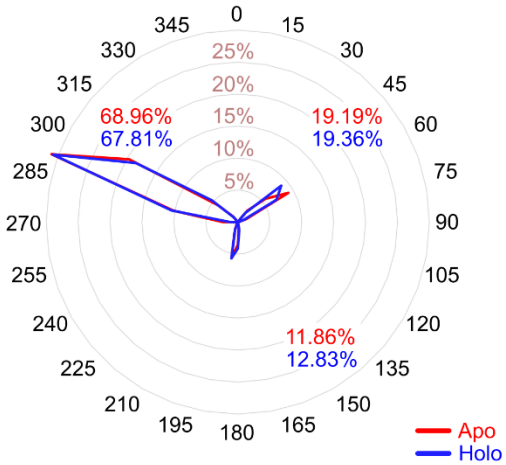
S2 G.



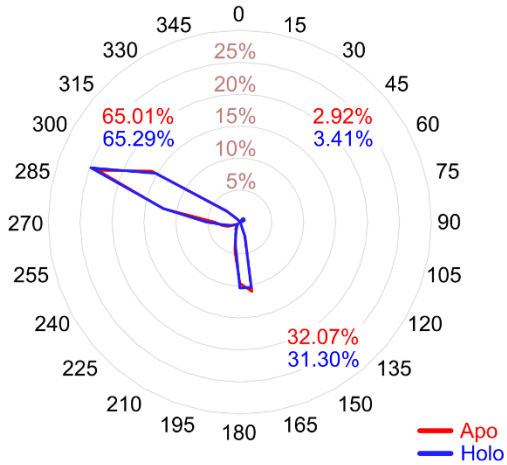
S2 H.



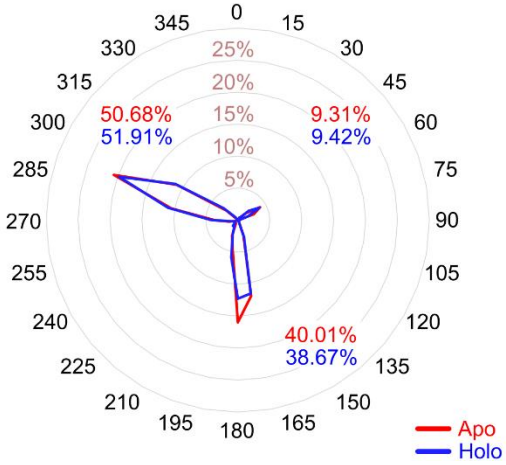
S2 I.



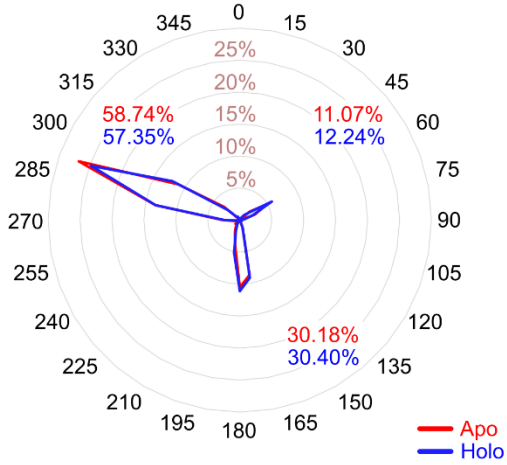
S2 J.



S2 K.



S2 L.



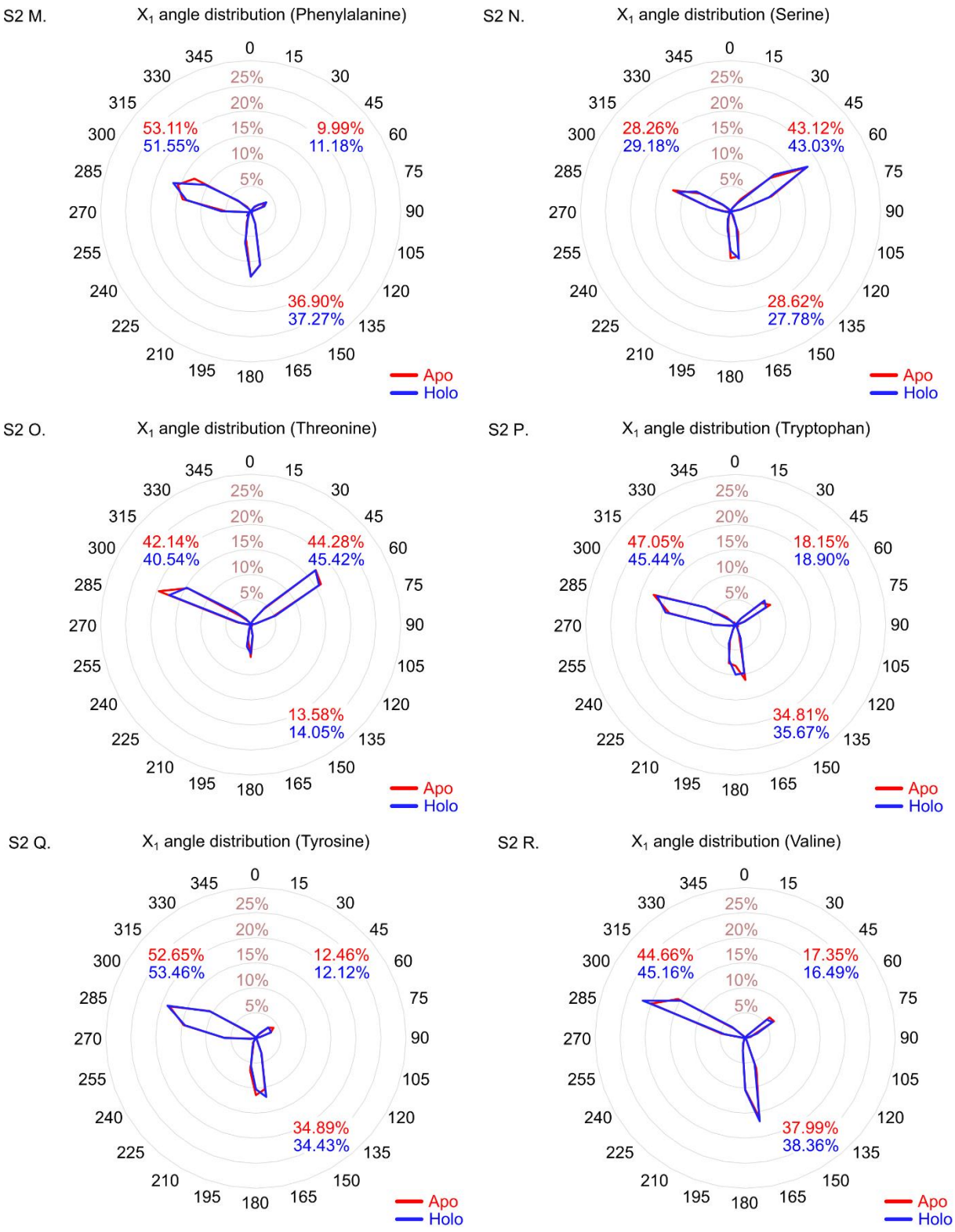


Figure A-2 (S2A–R.) Radar plots of χ_1 angle distributions. Distribution of χ_1 angles observed in unified binding site residues. Values were normalized on a per-family basis before radar binning such that each unique protein sequence is represented equally, regardless of family size. Data for: A) All UBS residues,

B) Arg, C) Asn, D) Asp, E) Cys, F) Gln, G) Glu, H) His, I) Ile, J) Leu, K) Lys, L) Met, M) Phe, N) Ser, O) Thr, P) Trp, Q) Tyr, R) Val.

Appendix B. Additional Figures for Protein-Protein Interfaces

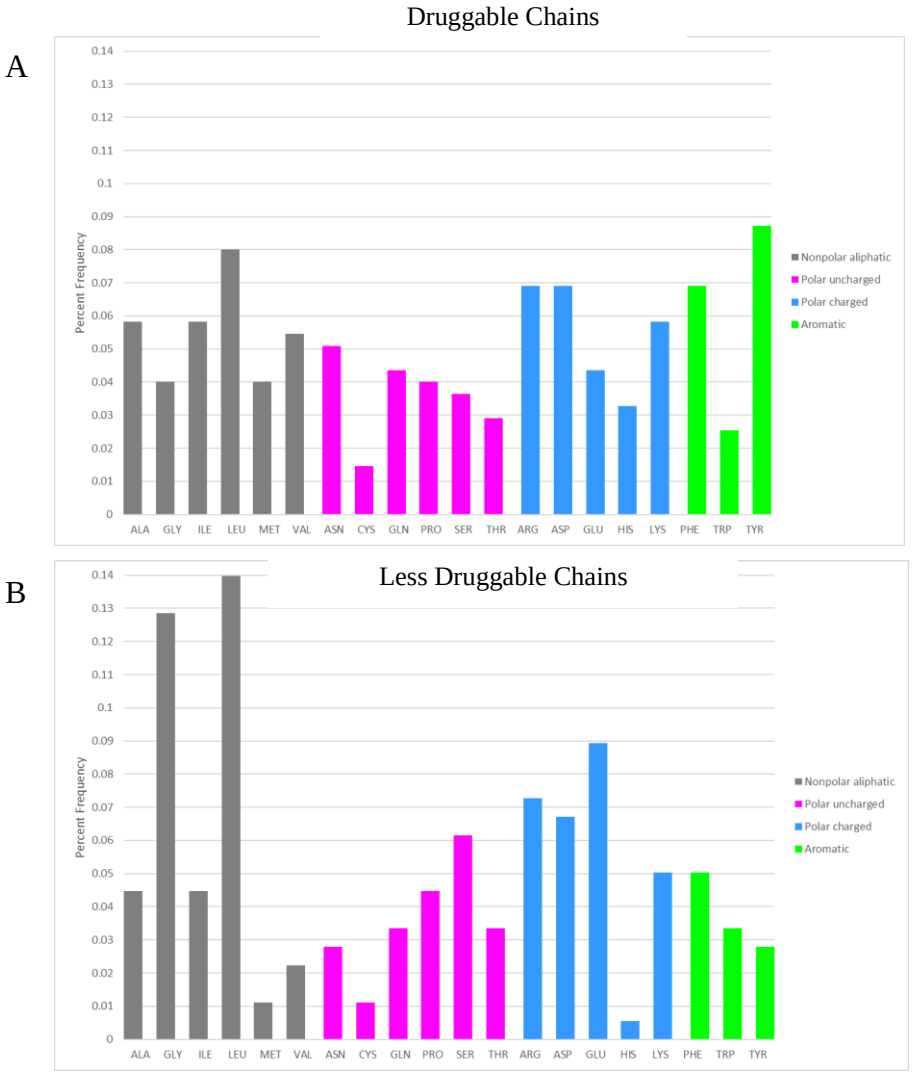
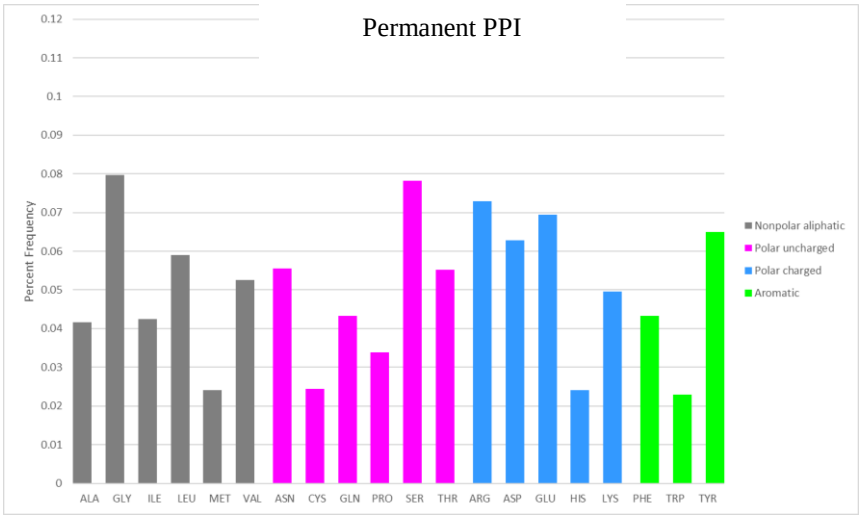
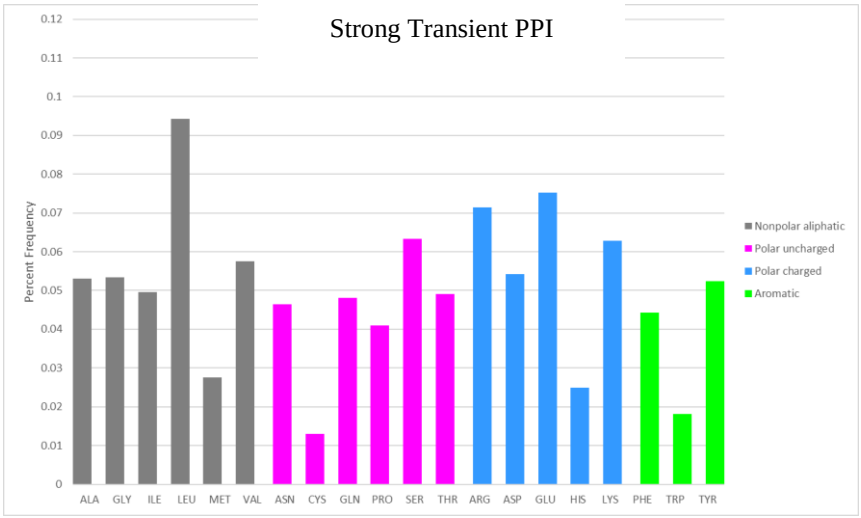


Figure B-1. Distributions of PPI contact residues for the 16 PPI complexes of the 2P2I set. Contact residues belonging to the: A) Druggable chains and B) Complementary chains

A


Permanent PPI
Percent Frequency
0.12
0.11
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
ALA GLY ILE LEU MET VAL ASN CYS GLN PRO SER THR ARG ASP GLU HIS LYS PHE TRP TYR
Nonpolar aliphatic
Polar uncharged
Polar charged
Aromatic

B


Strong Transient PPI
Percent Frequency
0.12
0.11
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
ALA GLY ILE LEU MET VAL ASN CYS GLN PRO SER THR ARG ASP GLU HIS LYS PHE TRP TYR
Nonpolar aliphatic
Polar uncharged
Polar charged
Aromatic

C

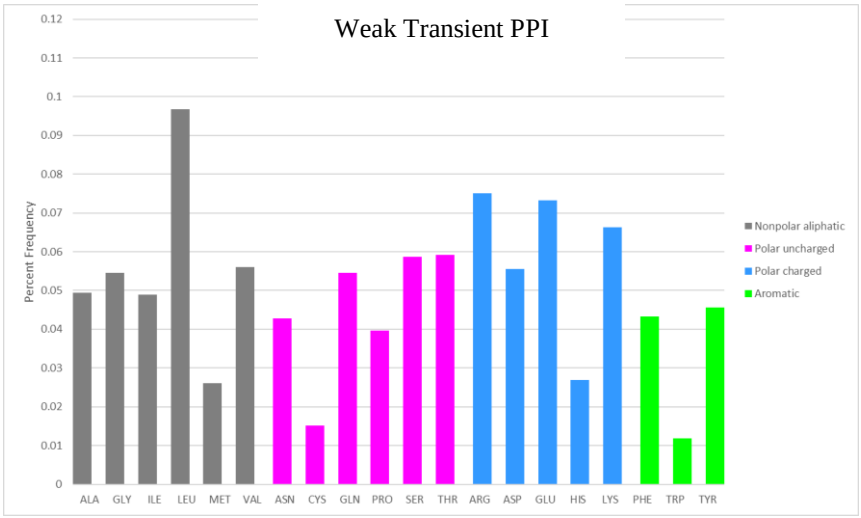

Weak Transient PPI
Percent Frequency
0.12
0.11
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
ALA GLY ILE LEU MET VAL ASN CYS GLN PRO SER THR ARG ASP GLU HIS LYS PHE TRP TYR
Nonpolar aliphatic
Polar uncharged
Polar charged
Aromatic

Figure B-2. Distributions of PPI contact residues for the complexes of the PDBbind set. Contact residues belonging to the: A) Permanent, B) Strong Transient, and C) Weak Transient complexes.

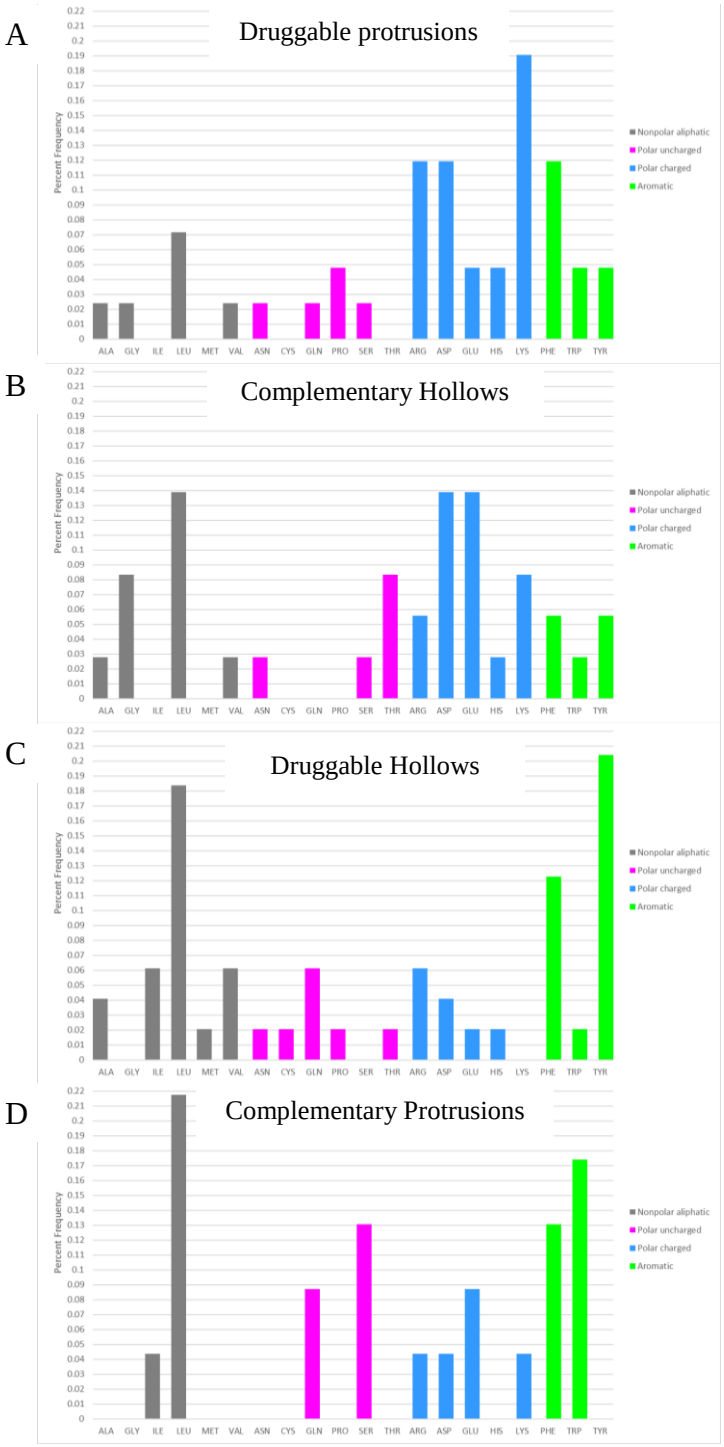


Figure B-3. Distributions of PPI contact residues representing the protrusions and hollows for the 16 PPI complexes of the P-P 2P2I set. Representative cluster residues belonging to: A) Druggable chain

protrusions, B) Complementary chain hollows, C) Complementary chain protrusions, D) Druggable chain hollows.

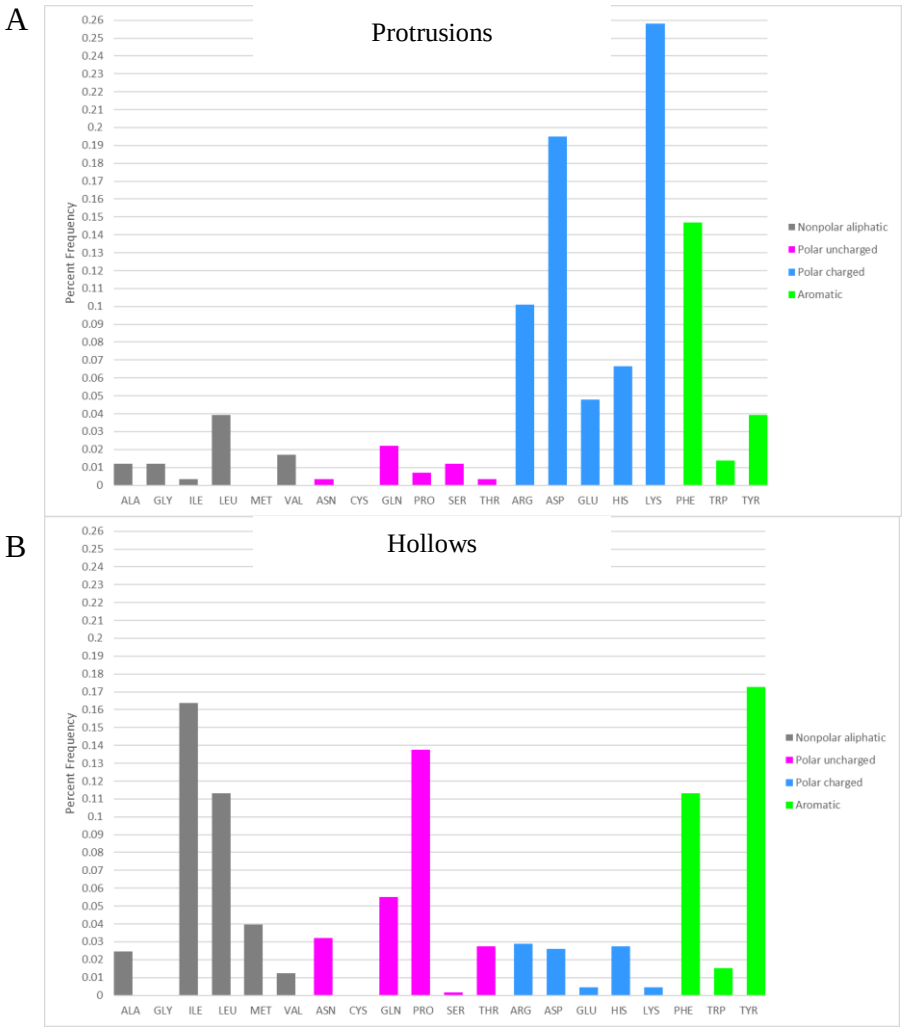
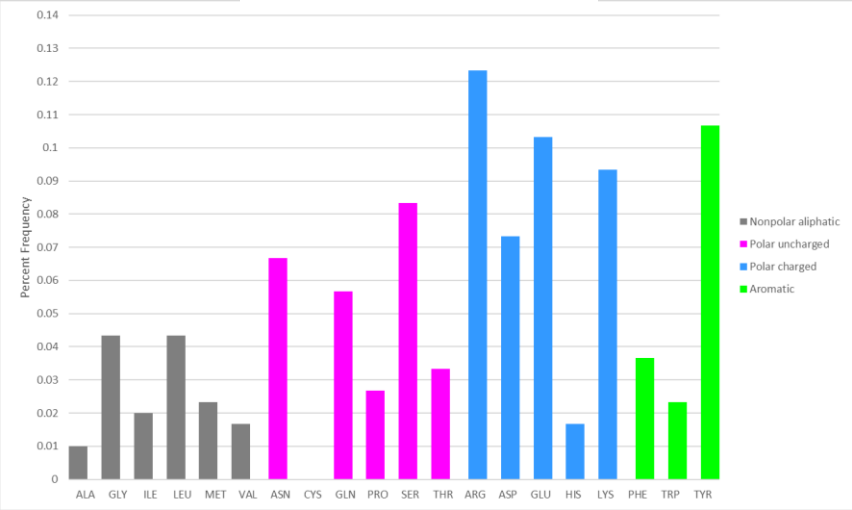


Figure B-4. Distributions of PPI contact residues representing the protrusions and hollows for the 204 ligand-bound PPI complexes of the P-L 2P2I set. Representative cluster residues belonging to the ligand-bound chain: A) protrusions, and B) hollows.

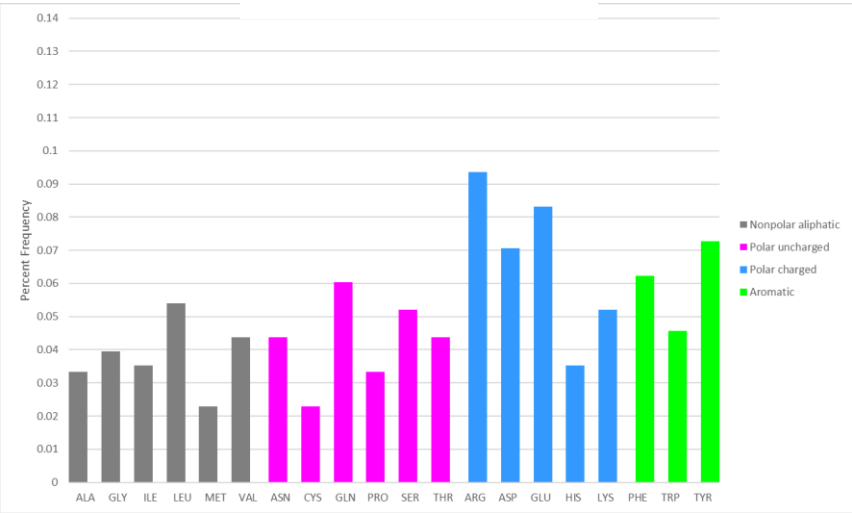
A

Permanent Protrusions



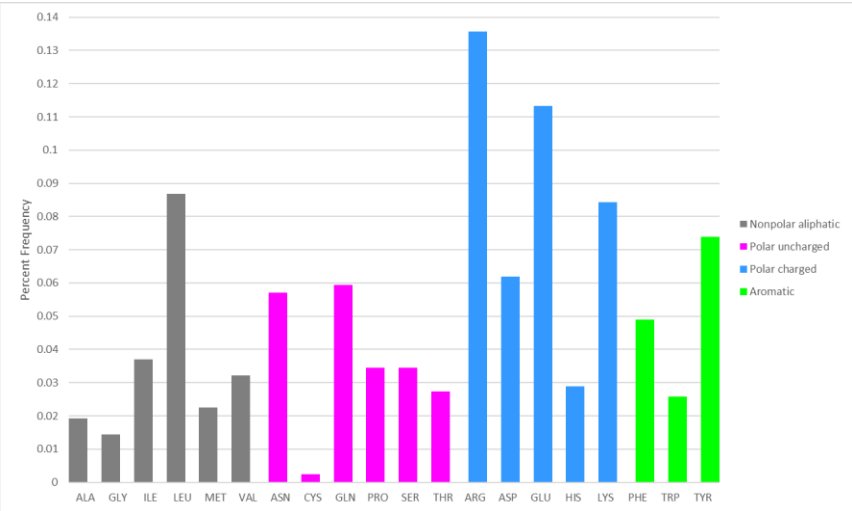
B

Permanent Hollows



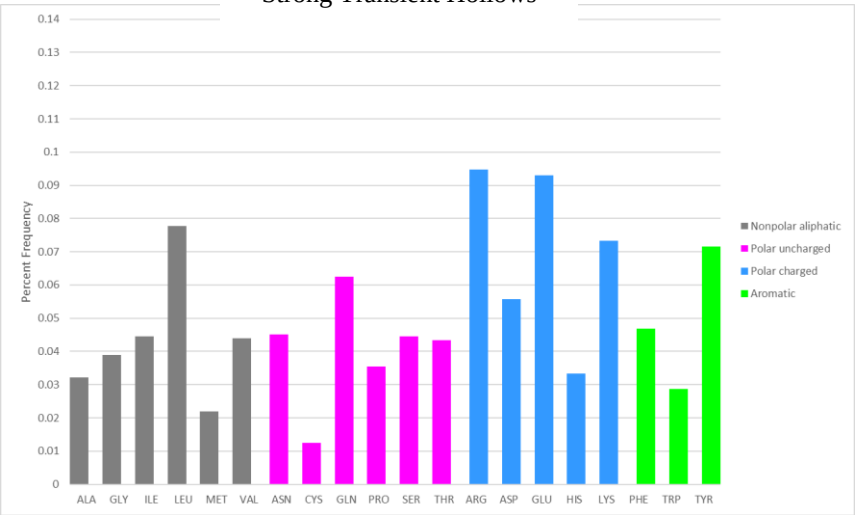
C

Strong Transient Protrusions



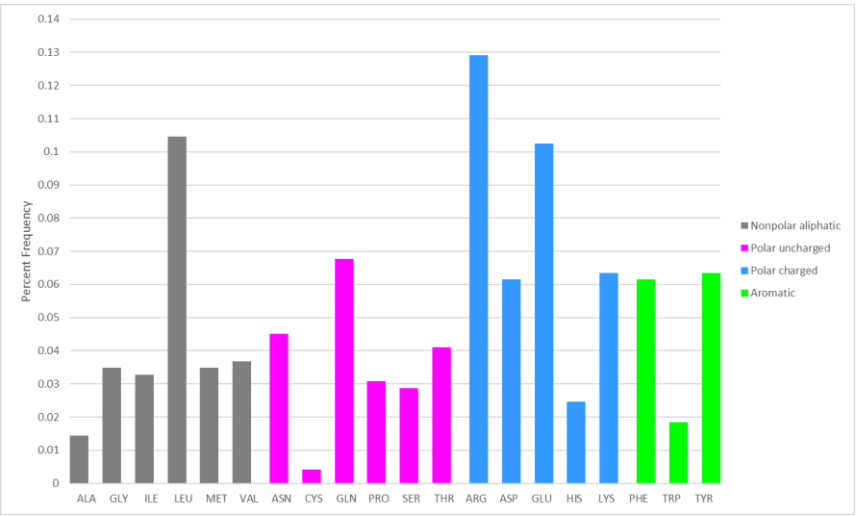
D

Strong Transient Hollows



E

Weak Transient Protrusions



F

Weak Transient Hollows

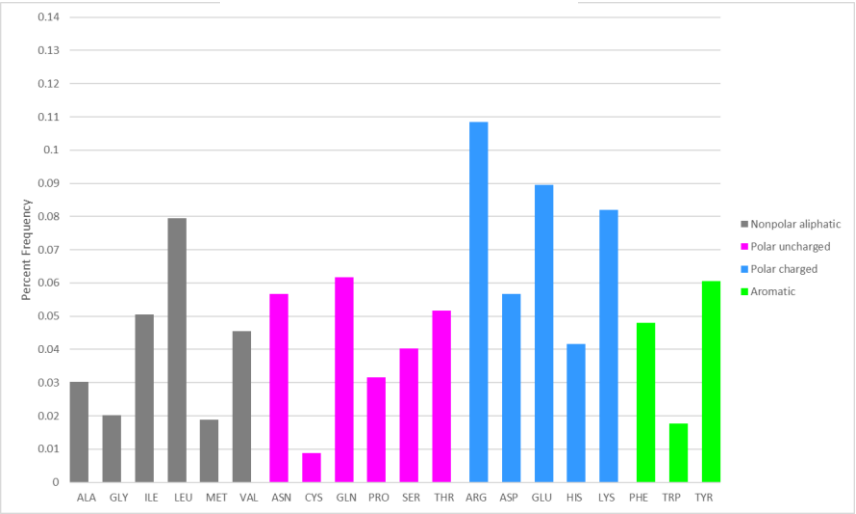


Figure B-5. Distributions of PPI contact residues representing the protrusions and hollows for the complexes of the PDBbind set. Contact residues belonging to the: A) permanent protrusions, B) permanent hollows, C) strong transient protrusions, D) strong transient hollows, E) weak transient protrusions, and F) weak transient hollows.

References

1. Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A., Binding MOAD (Mother Of All Databases). *Proteins* **2005,** *60* (3), 333-40.

2. Benson, M. L.; Smith, R. D.; Khazanov, N. A.; Dimcheff, B.; Beaver, J.; Dresslar, P.; Nerothin, J.; Carlson, H. A., Binding MOAD, a high-quality protein–ligand database. *Nucleic acids research* **2008,** *36* (suppl 1), D674-D678.

3. Ahmed, A.; Smith, R. D.; Clark, J. J.; Dunbar, J. B., Jr.; Carlson, H. A., Recent improvements to Binding MOAD: a resource for protein-ligand binding affinities and structures. *Nucleic Acids Res* **2015,** *43* (Database issue), D465-9.

4. Ma, B.; Shatsky, M.; Wolfson, H. J.; Nussinov, R., Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci* **2002,** *11* (2), 184-97.

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

6. Heringa, J.; Argos, P., Strain in protein structures as viewed through nonrotameric side chains: I. Their position and interaction. *Proteins-Structure Function and Genetics* **1999,** *37* (1), 30-43.

7. Heringa, J.; Argos, P., Strain in protein structures as viewed through nonrotameric side chains: II. Effects upon ligand binding. *Proteins-Structure Function and Genetics* **1999,** *37* (1), 44-55.

8. Freire, E., 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* **1999,** *96* (18), 10118-22.

9. Luque, I.; Freire, E., Structural stability of binding sites: consequences for binding affinity and allosteric effects. *Proteins* **2000,** *Suppl 4*, 63-71.

10. Shoichet, B. K.; Baase, W. A.; Kuroki, R.; Matthews, B. W., A relationship between protein stability and protein function. *Proc Natl Acad Sci U S A* **1995,** *92* (2), 452-6.

11. Jindal, G.; Warshel, A., Misunderstanding the preorganization concept can lead to confusions about the origin of enzyme catalysis. *Proteins* **2017,** *85* (12), 2157-2161.

12. Fischer, E., Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft* **1894,** *27* (3), 2985-2993.

13. Gohlke, H.; Klebe, G., Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angewandte Chemie International Edition* **2002,** *41* (15), 2644-2676.

14. Kunz, H., Emil Fischer—Unequalled Classicist, Master of Organic Chemistry Research, and Inspired Trailblazer of Biological Chemistry. *Angewandte Chemie International Edition* **2002,** *41* (23), 4439-4451.

15. Lipscomb, W. N., Linus Pauling 1901 - 1994. *Structure 2* (10), 991-992.

16. Pauling, L., Nature of Forces between Large Molecules of Biological Interest*. *Nature* **1948,** *161*, 707.

17. Koshland, D. E., Jr.; Ray, W. J., Jr.; Erwin, M. J., Protein structure and enzyme action. *Fed Proc* **1958,** *17* (4), 1145-50.

18. Carlson, H. A., Protein flexibility and drug design: how to hit a moving target. *Curr Opin Chem Biol* **2002,** *6* (4), 447-52.

19. Teague, S. J., Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov* **2003,** *2* (7), 527-41.

20. Carlson, H. A.; McCammon, J. A., Accommodating protein flexibility in computational drug design. *Mol Pharmacol* **2000,** *57* (2), 213-8.

21. Volkman, B. F.; Lipson, D.; Wemmer, D. E.; Kern, D., Two-state allosteric behavior in a single-domain signaling protein. *Science* **2001,** *291* (5512), 2429-33.

22. Finney, J. L.; Gellatly, B. J.; Golton, I. C.; Goodfellow, J., Solvent effects and polar interactions in the structural stability and dynamics of globular proteins. *Biophys J* **1980,** *32* (1), 17-33.

23. Gilson, M. K.; Zhou, H. X., Calculation of protein-ligand binding affinities. *Annu Rev Biophys Biomol Struct* **2007,** *36*, 21-42.

24. Gitlin, I.; Carbeck, J. D.; Whitesides, G. M., Why are proteins charged? Networks of charge-charge interactions in proteins measured by charge ladders and capillary electrophoresis. *Angew Chem Int Ed Engl* **2006,** *45* (19), 3022-60.

25. Frank, H. S.; Evans, M. W., Free Volume and Entropy in Condensed Systems III. Entropy in Binary Liquid Mixtures; Partial Molal Entropy in Dilute Solutions; Structure and Thermodynamics in Aqueous Electrolytes. *The Journal of Chemical Physics* **1945,** *13* (11), 507-532.

26. Sharp, K.; Nicholls, A.; Fine, R.; Honig, B., Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science* **1991,** *252* (5002), 106-109.

27. Olsson, T. S.; Williams, M. A.; Pitt, W. R.; Ladbury, J. E., The thermodynamics of protein-ligand interaction and solvation: insights for ligand design. *J Mol Biol* **2008,** *384* (4), 1002-17.

28. Chervenak, M. C.; Toone, E. J., A Direct Measure of the Contribution of Solvent Reorganization to the Enthalpy of Binding. *J. Am. Chem. Soc.* **1994,** *116* (23), 10533-10539.

29. Aqvist, J.; Medina, C.; Samuelsson, J. E., A new method for predicting binding affinity in computer-aided drug design. *Protein Eng* **1994,** *7* (3), 385-91.

30. Liang, J.; Edelsbrunner, H.; Woodward, C., Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci* **1998,** *7* (9), 1884-1897.

31. Scheiner, S., Comparison of CH...O, SH...O, Chalcogen, and Tetrel Bonds Formed by Neutral and Cationic Sulfur-Containing Compounds. *J Phys Chem A* **2015,** *119* (34), 9189-99.

32. Taylor, R.; Kennard, O., Crystallographic evidence for the existence of CH.cntdot..cntdot..cntdot.O, CH.cntdot..cntdot..cntdot.N and CH.cntdot..cntdot..cntdot.Cl hydrogen bonds. *J. Am. Chem. Soc.* **1982,** *104* (19), 5063-5070.

33. Pauling, L., *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry*. Cornell University Press: 1960.

34. Lafont, V.; Armstrong, A. A.; Ohtaka, H.; Kiso, Y.; Mario Amzel, L.; Freire, E., Compensating enthalpic and entropic changes hinder binding affinity optimization. *Chem Biol Drug Des* **2007,** *69* (6), 413-22.

35. Fonseca, T.; Ladanyi, B. M.; Hynes, J. T., Solvation free energies and solvent force constants. *J. Phys. Chem.* **1992,** *96* (10), 4085-4093.

36. Miyamoto, S.; Kollman, P. A., What determines the strength of noncovalent association of ligands to proteins in aqueous solution? *Proc Natl Acad Sci U S A* **1993,** *90* (18), 8402-6.

37. DeChancie, J.; Houk, K. N., The origins of femtomolar protein-ligand binding: hydrogen-bond cooperativity and desolvation energetics in the biotin-(strept)avidin binding site. *J Am Chem Soc* **2007,** *129* (17), 5419-29.

38. Hyre, D. E.; Le Trong, I.; Merritt, E. A.; Eccleston, J. F.; Green, N. M.; Stenkamp, R. E.; Stayton, P. S., Cooperative hydrogen bond interactions in the streptavidin-biotin system. *Protein Sci* **2006,** *15* (3), 459-67.

39. Xie, Z.-R.; Hwang, M.-J., Methods for Predicting Protein–Ligand Binding Sites. In *Molecular Modeling of Proteins*, Humana Press, New York, NY: 2015; pp 383-398.

40. Nooren, I. M.; Thornton, J. M., Diversity of protein-protein interactions. *EMBO J* **2003,** *22* (14), 3486-92.

41. Wells, J. A.; McClendon, C. L., Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **2007,** *450* (7172), 1001-9.

42. Vishwanath, S.; Sukhwal, A.; Sowdhamini, R.; Srinivasan, N., Specificity and stability of transient protein-protein interactions. *Curr Opin Struct Biol* **2017,** *44*, 77-86.

43. Jones, S.; Thornton, J. M., Protein-protein interactions: a review of protein dimer structures. *Prog Biophys Mol Biol* **1995,** *63* (1), 31-65.

44. Gowthaman, R.; Deeds, E. J.; Karanicolas, J., Structural properties of non-traditional drug targets present new challenges for virtual screening. *J Chem Inf Model* **2013,** *53* (8), 2073-81.

45. Fuller, J. C.; Burgoyne, N. J.; Jackson, R. M., Predicting druggable binding sites at the protein-protein interface. *Drug Discov Today* **2009,** *14* (3-4), 155-61.

46. Sledz, P.; Caflisch, A., Protein structure-based drug design: from docking to molecular dynamics. *Curr Opin Struct Biol* **2018,** *48*, 93-102.

47. Jones, S.; Thornton, J. M., Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **1996,** *93* (1), 13-20.

48. Keskin, O.; Ma, B.; Nussinov, R., Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol* **2005,** *345* (5), 1281-94.

49. Halperin, I.; Wolfson, H.; Nussinov, R., Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure* **2004,** *12* (6), 1027-38.

50. Keskin, O.; Gursoy, A.; Ma, B.; Nussinov, R., Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chem Rev* **2008,** *108* (4), 1225-44.

51. Janin, J.; Miller, S.; Chothia, C., Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol* **1988,** *204* (1), 155-64.

52. Lo Conte, L.; Chothia, C.; Janin, J., The atomic structure of protein-protein recognition sites. *J Mol Biol* **1999,** *285* (5), 2177-98.

53. Chakrabarti, P.; Janin, J., Dissecting protein-protein recognition sites. *Proteins* **2002,** *47* (3), 334-43.

54. Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J., Dissecting subunit interfaces in homodimeric proteins. *Proteins* **2003,** *53* (3), 708-19.

55. Basse, M. J.; Betzi, S.; Bourgeas, R.; Bouzidi, S.; Chetrit, B.; Hamon, V.; Morelli, X.; Roche, P., 2P2Idb: a structural database dedicated to orthosteric modulation of protein-protein interactions. *Nucleic Acids Res* **2013,** *41* (Database issue), D824-7.

56. Bourgeas, R.; Basse, M. J.; Morelli, X.; Roche, P., Atomic analysis of protein-protein interfaces with known inhibitors: the 2P2I database. *PLoS One* **2010,** *5* (3), e9598.

57. Basse, M. J.; Betzi, S.; Morelli, X.; Roche, P., 2P2Idb v2: update of a structural database dedicated to orthosteric modulation of protein-protein interactions. *Database (Oxford)* **2016,** *2016*.

58. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res* **2000,** *28* (1), 235-242.

59. Rose, P. W.; Prlic, A.; Altunkaya, A.; Bi, C.; Bradley, A. R.; Christie, C. H.; Costanzo, L. D.; Duarte, J. M.; Dutta, S.; Feng, Z.; Green, R. K.; Goodsell, D. S.; Hudson, B.; Kalro, T.; Lowe, R.; Peisach, E.; Randle, C.; Rose, A. S.; Shao, C.; Tao, Y. P.; Valasatava, Y.; Voigt, M.; Westbrook, J. D.; Woo, J.; Yang, H.; Young, J. Y.; Zardecki, C.; Berman, H. M.; Burley, S. K., The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res* **2017,** *45* (D1), D271-D281.

60. Wang, R.; Fang, X.; Lu, Y.; Wang, S., The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem* **2004,** *47* (12), 2977-80.

61. Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R., PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **2015,** *31* (3), 405-12.

62. Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R., Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *J Chem Inf Model* **2014,** *54* (6), 1700-16.

63. Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R., Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. *Acc Chem Res* **2017,** *50* (2), 302-309.

64. Dunbar, J. B., Jr.; Smith, R. D.; Damm-Ganamet, K. L.; Ahmed, A.; Esposito, E. X.; Delproposto, J.; Chinnaswamy, K.; Kang, Y. N.; Kubish, G.; Gestwicki, J. E.; Stuckey, J. A.;

Carlson, H. A., CSAR data set release 2012: ligands, affinities, complexes, and docking decoys. *J Chem Inf Model* **2013,** *53* (8), 1842-52.

65. Dunbar, J. B., Jr.; Smith, R. D.; Yang, C. Y.; Ung, P. M.; Lexa, K. W.; Khazanov, N. A.; Stuckey, J. A.; Wang, S.; Carlson, H. A., CSAR benchmark exercise of 2010: selection of the protein-ligand complexes. *J Chem Inf Model* **2011,** *51* (9), 2036-46.

66. Jérémy, D.; Guillaume, B.; Didier, R.; Esther, K., sc-PDB: a 3D-database of ligandable binding sites—10 years on. *Nucleic acids research* **2014,** *43* (D1), D399-404.

67. Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D., sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J Chem Inf Model* **2006,** *46* (2), 717-27.

68. Bhat, T. N.; Bourne, P.; Feng, Z.; Gilliland, G.; Jain, S.; Ravichandran, V.; Schneider, B.; Schneider, K.; Thanki, N.; Weissig, H.; Westbrook, J.; Berman, H. M., The PDB data uniformity project. *Nucleic Acids Res* **2001,** *29* (1), 214-8.

69. Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M., Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J Cheminform* **2014,** *6* (1), 12.

70. Yang, J.; Roy, A.; Zhang, Y., BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res* **2013,** *41* (Database issue), D1096-103.

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

72. Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J., BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* **2016,** *44* (D1), D1045-53.

73. Gathiaka, S.; Liu, S.; Chiu, M.; Yang, H.; Stuckey, J. A.; Kang, Y. N.; Delproposto, J.; Kubish, G.; Dunbar, J. B., Jr.; Carlson, H. A.; Burley, S. K.; Walters, W. P.; Amaro, R. E.; Feher, V. A.; Gilson, M. K., D3R grand challenge 2015: Evaluation of protein-ligand pose and affinity predictions. *J Comput Aided Mol Des* **2016,** *30* (9), 651-668.

74. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **2012,** *40* (Database issue), D1100-7.

75. Lu, J.; Carlson, H. A., ChemTreeMap: an interactive map of biochemical similarity in molecular datasets. *Bioinformatics* **2016,** *32* (23), 3584-3592.

76. Hendlich, M.; Bergner, A.; Günther, J.; Klebe, G., Relibase: Design and Development of a Database for Comprehensive Analysis of Protein–Ligand Interactions. *Journal of Molecular Biology* **2003,** *326* (2), 607-620.

77. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallographica Section B* **2016,** *72* (2), 171-179.

78. Roche, O.; Kiyama, R.; Brooks, C. L., 3rd, Ligand-protein database: linking protein-ligand complex structures to binding data. *J Med Chem* **2001,** *44* (22), 3592-8.

79. Block, P.; Sotriffer, C. A.; Dramburg, I.; Klebe, G., AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res* **2006,** *34* (Database issue), D522-6.

80. Konc, J.; Janezic, D., ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **2010,** *26* (9), 1160-8.

81. The UniProt, C., UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **2017,** *45* (D1), D158-D169.

82. Martin, A. C., Mapping PDB chains to UniProtKB entries. *Bioinformatics* **2005,** *21* (23), 4297-301.

83. Stajich, J. E.; Block, D.; Boulez, K.; Brenner, S. E.; Chervitz, S. A.; Dagdigian, C.; Fuellen, G.; Gilbert, J. G.; Korf, I.; Lapp, H.; Lehvaslaiho, H.; Matsalla, C.; Mungall, C. J.; Osborne, B. I.; Pocock, M. R.; Schattner, P.; Senger, M.; Stein, L. D.; Stupka, E.; Wilkinson, M. D.; Birney, E., The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **2002,** *12* (10), 1611-8.

84. Christianson, D. W., Structural biology of zinc. *Adv Protein Chem* **1991,** *42*, 281-355.

85. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L., BLAST+: architecture and applications. *BMC Bioinformatics* **2009,** *10*, 421.

86. Olson, S. A., EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. *Brief Bioinform* **2002,** *3* (1), 87-91.

87. Li, W.; Cowley, A.; Uludag, M.; Gur, T.; McWilliam, H.; Squizzato, S.; Park, Y. M.; Buso, N.; Lopez, R., The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res* **2015,** *43* (W1), W580-4.

88. Rose, A. S.; Hildebrand, P. W., NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res* **2015,** *43* (W1), W576-9.

89. Echols, N.; Milburn, D.; Gerstein, M., MolMovDB: analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Res* **2003,** *31* (1), 478-82.

90. Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S., A critical assessment of docking programs and scoring functions. *J Med Chem* **2006,** *49* (20), 5912-31.

91. Leach, A. R.; Shoichet, B. K.; Peishoff, C. E., Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J Med Chem* **2006,** *49* (20), 5851-5.

92. Murray, C. W.; Baxter, C. A.; Frenkel, A. D., The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *J Comput Aided Mol Des* **1999,** *13* (6), 547-62.

93. Zhao, Y.; Sanner, M. F., Protein-ligand docking with multiple flexible side chains. *J Comput Aided Mol Des* **2008,** *22* (9), 673-9.

94. May, A.; Zacharias, M., Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: evaluation on kinase inhibitor cross docking. *J Med Chem* **2008,** *51* (12), 3499-506.

95. Koska, J.; Spassov, V. Z.; Maynard, A. J.; Yan, L.; Austin, N.; Flook, P. K.; Venkatachalam, C. M., Fully automated molecular mechanics based induced fit protein-ligand docking method. *J Chem Inf Model* **2008,** *48* (10), 1965-73.

96. Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M., Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J Med Chem* **2004,** *47* (1), 45-55.

97. Gutteridge, A.; Thornton, J., Conformational change in substrate binding, catalysis and product release: an open and shut case? *FEBS Lett* **2004,** *567* (1), 67-73.

98. Gutteridge, A.; Thornton, J., Conformational changes observed in enzyme crystal structures upon substrate binding. *J Mol Biol* **2005,** *346* (1), 21-8.

99. Gunasekaran, K.; Nussinov, R., 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* **2007,** *365* (1), 257-73.

100. Brylinski, M.; Skolnick, J., What is the relationship between the global structures of apo and holo proteins? *Proteins* **2008,** *70* (2), 363-77.

101. Fradera, X.; De La Cruz, X.; Silva, C. H.; Gelpi, J. L.; Luque, F. J.; Orozco, M., Ligand-induced changes in the binding sites of proteins. *Bioinformatics* **2002,** *18* (7), 939-48.

102. Schnecke, V.; Kuhn, L. A., Database screening for HIV protease ligands: the influence of binding-site conformation and representation on ligand selectivity. *Proc Int Conf Intell Syst Mol Biol* **1999**, 242-51.

103. Schnecke, V.; Kuhn, L. A., Virtual screening with solvation and ligand-induced complementarity. *Perspectives in Drug Discovery and Design* **2000,** *20* (1), 171-190.

104. Zavodszky, M. I.; Kuhn, L. A., Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis. *Protein Sci* **2005,** *14* (4), 1104-14.

105. Zhao, S.; Goodsell, D. S.; Olson, A. J., Analysis of a data set of paired uncomplexed protein structures: new metrics for side-chain flexibility and model evaluation. *Proteins* **2001,** *43* (3), 271-9.

106. Najmanovich, R.; Kuttner, J.; Sobolev, V.; Edelman, M., Side-chain flexibility in proteins upon ligand binding. *Proteins-Structure Function and Genetics* **2000,** *39* (3), 261-268.

107. Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R., Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002,** *47* (4), 409-43.

108. Gaudreault, F.; Chartier, M.; Najmanovich, R., Side-chain rotamer changes upon ligand binding: common, crucial, correlate with entropy and rearrange hydrogen bonding. *Bioinformatics* **2012,** *28* (18), i423-i430.

109. Chang, D. T.; Yao, T. J.; Fan, C. Y.; Chiang, C. Y.; Bai, Y. H., AH-DB: collecting protein structure pairs before and after binding. *Nucleic Acids Res* **2012,** *40* (Database issue), D472-8.

110. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *J Mol Biol* **1990,** *215* (3), 403-10.

111. Damm, K. L.; Carlson, H. A., Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. *Biophys J* **2006,** *90* (12), 4558-73.

112. Hubbard, S.; Thornton, J., Naccess homepage: http://www.bioinf.manchester.ac.uk/naccess/ Accessed 04/01/2014.

113. *JMP*, Pro 11; SAS Institute INC.: Cary, NC, 1989-2018.

114. Carlson, H. A., Check your confidence: size really does matter. *J Chem Inf Model* **2013,** *53* (8), 1837-41.

115. Yang, C. Y.; Wang, R.; Wang, S., A systematic analysis of the effect of small-molecule binding on protein flexibility of the ligand-binding sites. *J Med Chem* **2005,** *48* (18), 5648-50.

116. Bogan, A. A.; Thorn, K. S., Anatomy of hot spots in protein interfaces. *J Mol Biol* **1998,** *280* (1), 1-9.

117. Marsh, J. A.; Teichmann, S. A., Relative solvent accessible surface area predicts protein conformational changes upon binding. *Structure* **2011,** *19* (6), 859-67.

118. Reynolds, C. H.; Tounge, B. A.; Bembenek, S. D., Ligand binding efficiency: trends, physical basis, and implications. *J Med Chem* **2008,** *51* (8), 2432-8.

119. Damm, K. L.; Carlson, H. A., Exploring experimental sources of multiple protein conformations in structure-based drug design. *J Am Chem Soc* **2007,** *129* (26), 8225-35.

120. Ghersi, D.; Sanchez, R., Beyond structural genomics: computational approaches for the identification of ligand binding sites in protein structures. *J Struct Funct Genomics* **2011,** *12* (2), 109-17.

121. Perot, S.; Sperandio, O.; Miteva, M. A.; Camproux, A. C.; Villoutreix, B. O., Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discov Today* **2010,** *15* (15-16), 656-67.

122. Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M., DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* **2008,** *36* (Database issue), D901-6.

123. Dessailly, B. H.; Lensink, M. F.; Orengo, C. A.; Wodak, S. J., LigASite—a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res* **2008,** *36* (suppl 1), D667-D673.

124. Laskowski, R. A., SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* **1995,** *13* (5), 323-30, 307-8.

125. Huang, B.; Schroeder, M., LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol* **2006,** *6*, 19.

126. Tan, K. P.; Varadarajan, R.; Madhusudhan, M. S., DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic Acids Res* **2011,** *39* (Web Server issue), W242-8.

127. Capra, J. A.; Laskowski, R. A.; Thornton, J. M.; Singh, M.; Funkhouser, T. A., Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput Biol* **2009,** *5* (12), e1000585.

128. Ravindranath, P. A.; Sanner, M. F., AutoSite: an automated approach for pseudo-ligands prediction-from ligand-binding sites identification to predicting key ligand atoms. *Bioinformatics* **2016,** *32* (20), 3142-3149.

129. Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W., Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem* **2007,** *50* (4), 726-41.

130. Wass, M. N.; Kelley, L. A.; Sternberg, M. J., 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res* **2010,** *38* (Web Server issue), W469-73.

131. Brylinski, M.; Skolnick, J., A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci U S A* **2008,** *105* (1), 129-34.

132. Lopez, G.; Valencia, A.; Tress, M. L., firestar--prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res* **2007,** *35* (Web Server issue), W573-7.

133. Zhang, Y., I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **2008,** *9*, 40.

134. McGuffin, L. J.; Roche, D. B., Automated tertiary structure prediction with accurate local model quality assessment using the IntFOLD-TS method. *Proteins* **2011,** *79 Suppl 10*, 137-46.

135. Kawabata, T., Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins* **2010,** *78* (5), 1195-211.

136. Le Guilloux, V.; Schmidtke, P.; Tuffery, P., Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **2009,** *10*, 168.

137. Hernandez, M.; Ghersi, D.; Sanchez, R., SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res* **2009,** *37* (Web Server issue), W413-6.

138. Laurie, A. T.; Jackson, R. M., Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **2005,** *21* (9), 1908-16.

139. Ngan, C. H.; Hall, D. R.; Zerbe, B.; Grove, L. E.; Kozakov, D.; Vajda, S., FTSite: high accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics* **2012,** *28* (2), 286-7.

140. Chen, P.; Huang, J. Z.; Gao, X., LigandRFs: random forest ensemble to identify ligand-binding residues from sequence information alone. *BMC Bioinformatics* **2014,** *15 Suppl 15*, S4.

141. Passerini, A.; Punta, M.; Ceroni, A.; Rost, B.; Frasconi, P., Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. *Proteins* **2006,** *65* (2), 305-16.

142. Shu, N.; Zhou, T.; Hovmoller, S., Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics* **2008,** *24* (6), 775-82.

143. Xie, Z. R.; Hwang, M. J., Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. *Bioinformatics* **2012,** *28* (12), 1579-85.

144. Xie, Z. R.; Liu, C. K.; Hsiao, F. C.; Yao, A.; Hwang, M. J., LISE: a server using ligand-interacting and site-enriched protein triangles for prediction of ligand-binding sites. *Nucleic Acids Res* **2013,** *41* (W1), W292-W296.

145. Mehio, W.; Kemp, G. J.; Taylor, P.; Walkinshaw, M. D., Identification of protein binding surfaces using surface triplet propensities. *Bioinformatics* **2010,** *26* (20), 2549-55.

146. Soga, S.; Shirai, H.; Kobori, M.; Hirayama, N., Use of amino acid composition to predict ligand-binding sites. *J Chem Inf Model* **2007,** *47* (2), 400-6.

147. Gutteridge, A.; Bartlett, G. J.; Thornton, J. M., Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J Mol Biol* **2003,** *330* (4), 719-34.

148. Kauffman, C.; Karypis, G., LIBRUS: combined machine learning and homology information for sequence-based ligand-binding residue prediction. *Bioinformatics* **2009,** *25* (23), 3099-107.

149. Komiyama, Y.; Banno, M.; Ueki, K.; Saad, G.; Shimizu, K., Automatic generation of bioinformatics tools for predicting protein-ligand binding sites. *Bioinformatics* **2016,** *32* (6), 901-7.

150. Huang, B., MetaPocket: a meta approach to improve protein ligand binding site prediction. *OMICS* **2009,** *13* (4), 325-30.

151. Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A., Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins* **2018,** *86 Suppl 1*, 7-15.

152. Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A., Critical assessment of methods of protein structure prediction (CASP)--round x. *Proteins* **2014,** *82 Suppl 2*, 1-6.

153. Cassarino, T. G.; Bordoli, L.; Schwede, T., Assessment of ligand binding site predictions in CASP10. *Proteins* **2014,** *82* (0 2), 154-163.

154. Haas, J.; Roth, S.; Arnold, K.; Kiefer, F.; Schmidt, T.; Bordoli, L.; Schwede, T., The Protein Model Portal--a comprehensive resource for protein structure and model information. *Database (Oxford)* **2013,** *2013*, bat031.

155. Hendlich, M.; Rippmann, F.; Barnickel, G., LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model* **1997,** *15* (6), 359-63, 389.

156. Connolly, M. L., Analytical Molecular-Surface Calculation. *Journal of Applied Crystallography* **1983,** *16* (Oct), 548-558.

157. Glaser, F.; Rosenberg, Y.; Kessel, A.; Pupko, T.; Ben-Tal, N., The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *Proteins* **2005,** *58* (3), 610-7.

158. Tan, K. P.; Nguyen, T. B.; Patel, S.; Varadarajan, R.; Madhusudhan, M. S., Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. *Nucleic Acids Res* **2013,** *41* (Web Server issue), W314-21.

159. Shrake, A.; Rupley, J. A., Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol* **1973,** *79* (2), 351-71.

160. Hubbard, T. J.; Blundell, T. L., Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng* **1987,** *1* (3), 159-71.

161. Morris, G. M.; Huey, R.; Olson, A. J., Using AutoDock for ligand-receptor docking. *Curr Protoc Bioinformatics* **2008,** *Chapter 8*, Unit 8 14.

162. Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S., A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* **2007,** *28* (6), 1145-52.

163. Fischer, J. D.; Mayer, C. E.; Soding, J., Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics* **2008,** *24* (5), 613-20.

164. Glaser, F.; Morris, R. J.; Najmanovich, R. J.; Laskowski, R. A.; Thornton, J. M., A method for localizing ligand binding pockets in protein structures. *Proteins* **2006,** *62* (2), 479-88.

165. Cruickshank, D. W. J., Remarks about protein structure precision. *Acta Crystallogr D* **1999,** *55* (3), 583-601.

166. Moult, J.; Fidelis, K.; Kryshtafovych, A.; Tramontano, A., Critical assessment of methods of protein structure prediction (CASP)--round IX. *Proteins* **2011,** *79 Suppl 10*, 1-5.

167. Chene, P., Drugs targeting protein-protein interactions. *ChemMedChem* **2006,** *1* (4), 400-11.

168. Fry, D. C., Protein-protein interactions as targets for small molecule drug discovery. *Biopolymers* **2006,** *84* (6), 535-52.

169. La, D.; Kong, M.; Hoffman, W.; Choi, Y. I.; Kihara, D., Predicting Permanent and Transient Protein-Protein Interfaces. *Proteins* **2013,** *81* (5), 805-818.

170. Wanner, J.; Fry, D. C.; Peng, Z.; Roberts, J., Druggability assessment of protein-protein interfaces. *Future Med Chem* **2011,** *3* (16), 2021-38.

171. Fiamegos, Y. C.; Kastritis, P. L.; Exarchou, V.; Han, H.; Bonvin, A. M.; Vervoort, J.; Lewis, K.; Hamblin, M. R.; Tegos, G. P., Antimicrobial and efflux pump inhibitory activity of caffeoylquinic acids from Artemisia absinthium against gram-positive pathogenic bacteria. *PLoS One* **2011,** *6* (4), e18127.

172. Sakono, M.; Zako, T., Amyloid oligomers: formation and toxicity of Abeta oligomers. *FEBS J* **2010,** *277* (6), 1348-58.

173. Davis, A. M.; Teague, S. J., Hydrogen Bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis. *Angew Chem Int Ed Engl* **1999,** *38* (6), 736-749.

174. Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S., Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* **2007,** *25* (1), 71-5.

175. Fry, D. C., Targeting protein-protein interactions for drug discovery. *Methods Mol Biol* **2015,** *1278*, 93-106.

176. Fry, D. C., Small-molecule inhibitors of protein-protein interactions: how to mimic a protein partner. *Curr Pharm Des* **2012,** *18* (30), 4679-84.

177. Argos, P., An investigation of protein subunit and domain interfaces. *Protein Eng* **1988,** *2* (2), 101-13.

178. Janin, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M. J.; Vajda, S.; Vakser, I.; Wodak, S. J.; Critical Assessment of, P. I., CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* **2003,** *52* (1), 2-9.

179. Moult, J.; Hubbard, T.; Bryant, S. H.; Fidelis, K.; Pedersen, J. T., Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins* **1997,** *Suppl 1*, 2-6.

180. Lensink, M. F.; Velankar, S.; Wodak, S. J., Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. *Proteins* **2017,** *85* (3), 359-377.

181. Douguet, D.; Chen, H. C.; Tovchigrechko, A.; Vakser, I. A., DOCKGROUND resource for studying protein-protein interfaces. *Bioinformatics* **2006,** *22* (21), 2612-8.

182. Higueruelo, A. P.; Schreyer, A.; Bickerton, G. R.; Pitt, W. R.; Groom, C. R.; Blundell, T. L., Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. *Chem Biol Drug Des* **2009,** *74* (5), 457-67.

183. Higueruelo, A. P.; Jubb, H.; Blundell, T. L., TIMBAL v2: update of a database holding small molecules modulating protein-protein interactions. *Database (Oxford)* **2013,** *2013*, bat039.

184. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magarinos, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R., The ChEMBL database in 2017. *Nucleic Acids Res* **2017,** *45* (D1), D945-D954.

185. Labbe, C. M.; Laconde, G.; Kuenemann, M. A.; Villoutreix, B. O.; Sperandio, O., iPPI-DB: a manually curated and interactive database of small non-peptide inhibitors of protein-protein interactions. *Drug Discov Today* **2013,** *18* (19-20), 958-68.

186. Labbe, C. M.; Kuenemann, M. A.; Zarzycka, B.; Vriend, G.; Nicolaes, G. A.; Lagorce, D.; Miteva, M. A.; Villoutreix, B. O.; Sperandio, O., iPPI-DB: an online database of modulators of protein-protein interactions. *Nucleic Acids Res* **2016,** *44* (D1), D542-7.

187. Reynolds, C.; Damerell, D.; Jones, S., ProtorP: a protein-protein interaction analysis server. *Bioinformatics* **2009,** *25* (3), 413-4.

188. Negi, S. S.; Schein, C. H.; Oezguen, N.; Power, T. D.; Braun, W., InterProSurf: a web server for predicting interacting sites on protein surfaces. *Bioinformatics* **2007,** *23* (24), 3397-9.

189. Laskowski, R. A., PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res* **2001,** *29* (1), 221-2.

190. Laskowski, R. A.; Jablonska, J.; Pravda, L.; Varekova, R. S.; Thornton, J. M., PDBsum: Structural summaries of PDB entries. *Protein Sci* **2018,** *27* (1), 129-134.

191. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G., ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* **2012,** *52* (7), 1757-68.

192. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K., Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* **2012,** *55* (14), 6582-94.

193. McGann, M., FRED and HYBRID docking performance on standardized datasets. *J Comput Aided Mol Des* **2012,** *26* (8), 897-906.

194. Khazanov, N. A.; Damm-Ganamet, K. L.; Quang, D. X.; Carlson, H. A., Overcoming sequence misalignments with weighted structural superposition. *Proteins* **2012,** *80* (11), 2523-35.

195. *MATLAB and Statistics Toolbox Release 2015a*, 2015a; The MathWorks, Inc.: Natick, Massachusetts, United States of America, 2015.

196. Schrodinger, LLC, The PyMOL Molecular Graphics System, Version 1.8. 2015.

197. Nooren, I. M. A.; Thornton, J. M., Structural Characterisation and Functional Significance of Transient Protein–Protein Interactions. *Journal of Molecular Biology* **2003,** *325* (5), 991-1018.