**Title**
Applications of a Surface-Based Protein Binding Site Comparison Methodology

**Author**
Spitzer, Russell Alexander

**Publication Date**
2013

Applications of a Surface-Based Protein Binding Site Comparison Methodology

by

Russell Alexander Spitzer

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

*To my wife Marguerite . . .*
*she always believes in me*
*and teaches me more than*
*I could have ever imagined*

# Acknowledgments

All research is a collaborative process, and this dissertation is no exception. The work presented here would have been impossible without the direct and integral help of the Jain Lab. My two labmates, Emmanuael Yera and Rocco Varela, helped me through many of the difficult academic and personal challenges that I encountered during my doctoral career. I can only hope that all students can find peers as intelligent and supportive as mine. Ann Cleves was always ready to lend a helping hand or provide a fresh perspective on my work. She often was able to discern interesting results even when I had lost faith. Finally, Ajay Jain, my advisor, not only provided a solid foundation on which to begin my research, but also was a constant motivator and collaborator.

I would also be remiss, if I did not mention my orals and thesis committee members Hana El-Samed, Hao Li, Patricia Babbit and Andre Šali. Our meetings were always helpful in setting new goals and making sure I did not get off track. Thanks as well to the NIH and UCSF, whose financial and institutional support made it possible for me to pursue this degree.

I would also like to thank my family, who have at many times sat through my practice talks and listened to me ramble at lengths about my work. They have been extremely supportive and I know that without them I would never have even considered applying to graduate school. Special thanks needs to go to my wife Marguerite Sheffer, whose patience and strength were invaluable.

## Previous Publications

Chapters 2,3 have already been published as articles in peer-reviewed journals. The publication details appear at the beginning of each chapter. Chapter 4 has been submitted and is awaiting review.

## Statement from Ajay Jain, thesis advisor, on chapter co-authors:

Chapter 2 was adapted from the following paper: R. Spitzer, A. E. Cleves, and A. N. Jain. Surface-based protein binding pocket similarity. Proteins: Struct., Funct., Bioinf., 79(9):274663, 2011. Russell designed the primary experiments, implemented the relevant algorithms, performed essentially all computations, and made the majority of the analyses. Contributions came from all authors on control experiments and some aspects of data analysis.

Chapter 3 was adapted from: R. Spitzer and A. N. Jain. Surflex-dock: Docking benchmarks and real-world application. J. Comput.-Aided Mol. Des., 26(6):687699, 2012. Russell performed the experiments relating to multi-structure docking as well as running several controls. He also performed the experiments investigating the construction of designed decoy sets. The analysis and authorship were performed in concert with myself.

Chapter 4 was largely adapted from a paper, to be published: R. Spitzer, R. Varela, A.E. Cleves, and A.N. Jain. Protein Function Annotation By Local Binding Site Surface Similarity. Russell designed the primary experiments, implemented the relevant algorithms, performed all computations, and made the majority of the analyses. Contributions came from all authors on aspects of data analysis, data preparation, and literature review.

# Abstract

Applications of a Surface-Based
Protein Binding Site Comparison Methodology

Russell Alexander Spitzer

Protein similarity has been used for the annotation and classification of proteins when the structure of the protein is available. Protein similarity comparisons may be made on a local or global basis and may consider sequence information and differing levels of structural information. This dissertation details the method Surflex PSIM, a local 3D method that compares the surfaces of protein binding sites.

PSIM is a local 3D method that compares protein binding site surfaces in full atomic detail. The approach is based on the morphological similarity method (Surflex-Sim) which has been widely applied for global comparison of small molecules. This methodology has the ability to determine the differences between very similar proteins with different ligand binding specificity and the ability to correctly align extremely divergent proteins with only a small region of similarity. PSIM performed well on known standards for binding site comparisons.

In a docking benchmark study, PSIM was used to assist in multi-structure docking protocols. In these protocols, proper selection of target structures can reduce time required for screening and increase accuracy. Selection of a minimal representative set of docking target conformations was performed automatically using PSIM. Several docking targets, for which unsatisfactory results had been obtained used a single-structure protocol, yielded substantial improvements using the PSIM-aided multi-structure docking protocol.

Further development of an automated binding-site detection algorithm allowed for

PSIM to be used as screening tool for annotating proteins with unknown function. A dataset was created of proteins whose function was determined after their crystallization. PSIM was able to automatically detect binding sites on a majority of these proteins and successfully match them to proteins that were present in the PDB at the time of crystallization that have the same function. PSIM was further used to explore possible functions for several proteins whose function is still unknown.

The main contribution of this dissertation is a fast and accurate method for the comparison of protein binding sites agnostic of sequence information. This methodology has applications in the analysis of ligand specificity analysis and the annotation of proteins with unknown function.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Background

The specific binding of ligands to proteins is a fundamental aspect of the molecular machinery of organisms. Although the study of the interaction of ligands with protein binding sites is conceptually straightforward, characterization at a level that supports prediction of protein function or the identification of potential ligands remains challenging. The Protein Structure Initiative contains hundreds of currently unannotated protein structures, many without bound ligands, whose functions are unknown.[1] Even among proteins whose functions are known, such as human kinases, it is difficult to develop ligands which will specifically target a single kinase. These challenges have led to the development of several methodologies for computing the similarity between proteins.

## 1.2 Alternative Methods

In general, these protein comparisons algorithms can be grouped based upon their scope, what region of the protein is compared, and their underlying metric for measuring similarity. While not exhaustive, we will breifly discuss various approaches currently being used to measure similarity between protein structures.

Global similarity metrics attempt to find the largest matching region between two proteins. The goal of these methods is to find a similar fold topology or general structural correspondence between two proteins. These methodologies are generally unable to characterize small but significant regions of similarity when those regions are surrounded by divergent structures. Similarity calculated by these methods is usually dominated by similar secondary structure and scaffolding. Examples of this type of algorithm are the DALI[2] and CE[3] methods.

Motif based methods, such as SeqFeature,[4] look for common residues between the proteins being compared thus limiting them to comparing proteins with similar sequences. Other methods focus on comparing bit-strings of information which describe a protein binding site. These "fingerprint" methods include FuzCav[5] and PocketMatch[6] and are very fast, but generally insensitive to subtle conformational shifts.

Geometric methodologies utilize the structural geometry in a protein binding site to make comparisons and alignments. This can be accomplished by subdividing the binding site into triangles and comparing those, as with SiteBase,[7] or by comparing low-order spherical harmonics such as with SurfNet.[8] These methodologies tend to be more computationally intensive because of their full usage of structural information but are ideal for comparing small variations in binding sites. The lack of any connection to underlying sequence also makes these methods well suited for comparing extremely divergent proteins.

As an alternative to global similarity methods, local approaches define a region of comparison or scope and then check for similarity only within that defined region. This scope makes local approaches ideal for comparing smaller structural motifs or protein binding sites. Unlike global methods, these methods will ignore divergence outside of the scoped region, allowing for the correct comparison of proteins with differing folds but similar binding sites.

## 1.3 Surflex-PSIM

Surflex-PSIM was developed as a geometric method for local protein comparison. Unlike the global, motif-based, and fingerprint methods listed above, Surflex-PSIM focuses solely on the comparison of the surfaces of protein binding sites. Protein similarity is considered from the perspective of a ligand, investigating only those moieties which might interact with a bound ligand. Surflex-PSIM measures the differences between the composition and geometry of the local surfaces of protein binding sites. Key to this analysis of the binding site environment, is the realization that protein/ligand interactions are not defined by protein scaffolding. The surface features and geometry of the protein-binding site dictate the size, pose, and chemical features of potential ligands. A metric that successfully accounts for these variables will allow us to make predictions of protein ligand interactions for purposes such as specificity profiles and functional annotation of uncharacterized proteins.

An example of the application of Surflex-PSIM can be seen in Figure 1.1. Surflex-PSIM was used to align 8 crystal structures from the Enolase Superfamily designated by the Structure Function Linkage Database.[9] Of these structures, 4 were members of the Mandelate racemase subgroup part and 4 were members of the Muconate cycloisomerase subgroup. Using Surflex-PSIM to produce a common alignment yielded the tree seen in the upper right of Figure 1.1. PSIM assembles this tree by greedily connecting each structure to the structure with which it is most similar. This automatic methodology segregated the two subgroups while still recognizing the common structural motifs between all 8 structures. In the upper left and lower right, two examples from each subgroup are presented in alignment, showing that Surflex-PSIM correctly aligned the structures. In the left, we see that the alignment between the subgroups also correctly positions the binding sites and co-localizes the magnesium cofactor present in the structures. PSIM finds a common alignment between members of a superfamily while also automatically dividing the structures into functionally annotated subgroups.

## 1.4 Synopsis

In this work we document several applications of Surflex-PSIM. Chapter 2 provides an introduction to and description of the method. This work, demonstrates that Surflex-PSIM is sensitive enough to detect the subtle conformational shifts caused by ligand binding, while still being able to automatically detect 3-dimensional motifs in highly diverse proteins. Also, Surflex-PSIM is shown to be able to discriminate between human kinases which are inhibited by the same ligand and those which are not. Chapter 3 describes the ability to detect the subtle conformational shifts of a single protein used in a docking-benchmark study. This application shows a significant benefit to using multiple-aligned structures as input to a standard screening procedure. Chapter 4, shows that Surflex-PSIM is capable of annotating un-characterized proteins. Performing well on a time-segregated data-set and positing new functions for currently un-annotated proteins.

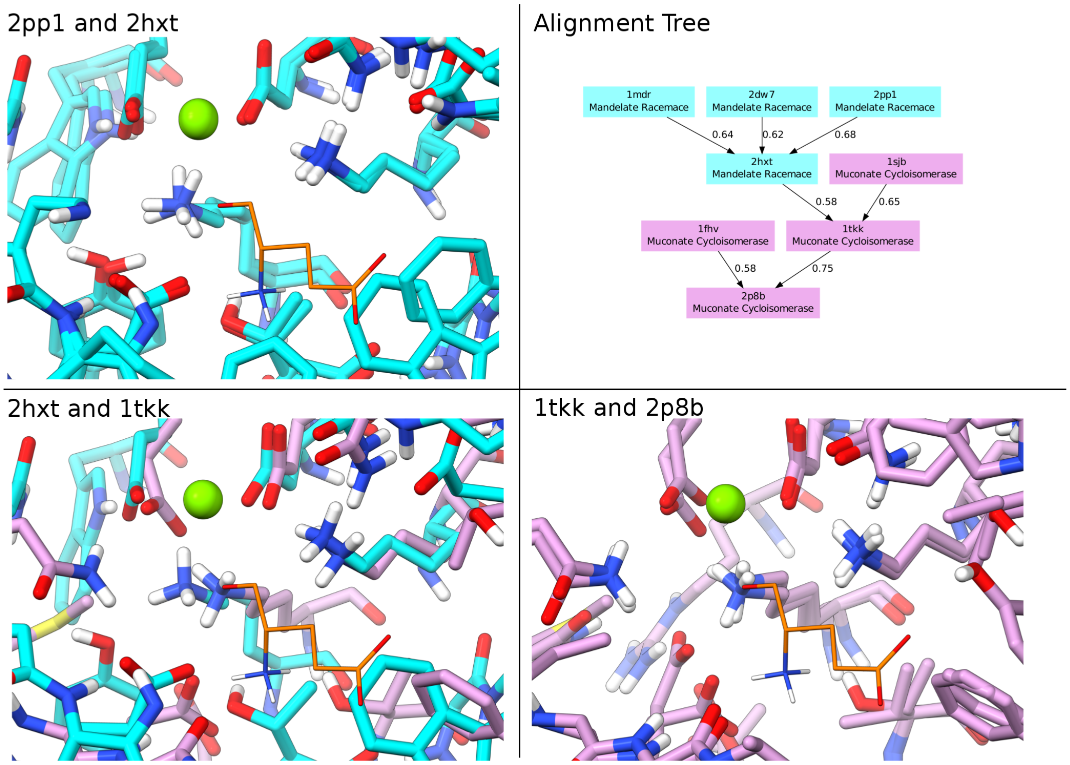


Figure 1.1: In the upper right, an tree showing the automatic alignment performed by Surflex-PSIM. Nodes are colored based on their Enolase subgroup membership (Mandelate Racemace in blue, Muconate Cycloisomerase in pink). Edges are marked with the raw similarity scores. In the upper left, the alignment of 2pp1 and 2hxt, both members of the Mandelate Racemace subgroup. In the lower left, the alignment of 2hxt and 1tkk, members of the Muconate Cycloisomerase and Mandelate Racemace subgroups respectively. In the lower right, the alignment of two Muconate Cycloisomerase structures 1tkk and 2p8b. Glutamic acid is shown in orange in all of the alignments to provide orientation within the binding site.

# Chapter 2

# Surface-based Protein Binding Pocket Similarity

Spitzer, Russell, Ann E. Cleves, and Ajay N. Jain.
"Surface-based protein binding pocket similarity."
*Proteins: Structure, Function, and Bioinformatics*
79.9 (2011): 2746-2763.

## 2.1 Abstract

Protein similarity comparisons may be made on a local or global basis and may consider sequence information or differing levels of structural information. We present a local three-dimensional method that compares protein binding site surfaces in full atomic detail. The approach is based on a morphological similarity method which has been widely applied for global comparison of small molecules. We apply the method to all-by-all comparisons two sets of human protein kinases, a very diverse set of ATP-bound proteins from multiple species, and three heterogeneous benchmark protein binding site data sets. Cases of disagreement between sequence-based similarity and binding site similarity yield informative examples. Where sequence similarity is very low, high pocket similarity can reliably identify important binding motifs. Where sequence similarity is very high, significant differences in pocket similarity are related to ligand binding specificity and similarity. Local protein binding pocket similarity provides qualitatively complementary information to other approaches, and it can yield quantitative information in support of functional annotation.

## 2.2 Introduction

Comparisons of small molecules based on surface characteristics including both shape and polarity have been shown to yield separable similarities for pairs of molecules that bind the same protein sites from those that do not.[10] Variations of the approach have been used in virtual screening[11] and for identifying molecular superimpositions for use in constructing binding site models for affinity prediction based on the structures and activities of small molecules targeting a single protein cavity.[12,13] The approach is based on defining molecular observers around two aligned molecules and comparing what they see in terms of distances to the molecular surfaces and the polarity of those surfaces (efficient solutions exist for the problem of identifying the optimal alignment and conformation of one molecule onto another). Recently, we have extended the approach to comparisons of concave surfaces such as protein binding pockets.[14]

This extension (from global/convex to local/concave) required two significant additions to the similarity computation: (1) a method to define the spatial scope of the desired comparison; and (2) methods to avoid degeneracies such as solvent accessibility that are present in the concavity-comparison case and do not arise otherwise. The software implementing the algorithms for pocket construction and ligand activity prediction constitute a new module within the Surflex platform, called Surflex-PSIM (Protein Similarity). In this article, we present the details of the similarity computation and optimization algorithm and application of the approach to three sets of related protein structures of varying diversity, two consisting of sets of protein kinases and the other consisting of evolutionarily-divergent ATP binding proteins. We compare the results obtained from these local structural comparisons with those obtained based on sequence comparisons. Although the sequence-based approach is clearly able to identify ancestral relationships between proteins, the surface-based approach offers complementary information allowing for more subtle distinctions that relate to protein function, especially those functions related to noncovalent ligand binding within protein cavities. We also present direct comparisons with other methods on three sets of heterogeneous protein structures reported by Kahraman et al.,[15] Hoffman et al.,[8] and Yeturu and Chandra.[6]

The first kinase set consisted of 45 structures of the MAP-kinase family EC 2.7.11.24. The second kinase set contained 183 protein structures, corresponding to 26 different human kinases, for which ligand binding data were available, as in Fabien et al.[16] Our quantitative pocket comparisons mirrored those found by Kinnings and Jackson.[7] This separation efficiency was superior to using sequence-based methodologies for several of our target ligands. The kinases represented a group of proteins that diverged relatively late in evolutionary time (based on amino acid similarity), whose basic function is the same but with differences in specificity and in regulation. Differences among these proteins included very subtle alterations of ligand binding pockets, and even the more substantial changes still preserved overall protein architecture and clear global sequence similarities.

ATP has been a metabolically important small molecule for the entire duration of evolution, so those proteins that make use of its properties have evolved a number of different structural mechanisms for doing so over a very long time span. The ATP-bound set contained 267 protein structures, corresponding to 120 annotated Enzyme Commission numbers, and many different families of proteins (including 6 helicases, 12 ligases, 39 metabolic kinases, 28 protein kinases, 13 polymerases, 5 protein folding chaperones, 29 tRNA synthetases, 8 transcription factors, and 21 transporters) from 78 species. While all of these proteins bind ATP, there is a great deal of variety in both the sequences of the proteins as well as the physical motifs used for binding. Comparisons using pocket similarity here identified cases as in the kinase set with high sequence similarity and a range of pocket similarities. More interestingly, a number of cases were identified with very high pocket similarity where the sequence similarity was so minimal as to be undetectable using standard sequence alignment methods.

Analysis of the data sets of Kahraman (100 proteins binding nine different ligand types), Hoffman (100 proteins binding 10 different ligands), and Yeturu (26 protein structures, with 51 ligand binding sites for four ligands) illuminated the differences between the Surflex-PSIM approach and other methods. The Kahraman set includes protein cavities of widely varying volume, with the Hoffman set designed specifically to avoid such gross heterogeneity. In these cases, Surflex-PSIM performed statistically indistinguishably from the best of previously reported methods. In contrast to those methods, the PSIM approach showed only a limited correlation with pocket volume differences, while showing a significant relationship between pocket similarity and cognate ligand similarity. On the Yeturu set, whose focus was on the binding sites of four ligands, where each binding site was represented by highly similar variants, the Surflex-PSIM approach yielded a perfect segregation of the sites by cognate ligand.

## 2.3 Methods and Data

Data sets comprising human protein kinases and evolutionarily diverse ATP-bound proteins were the subject of our primary analyses of proteins with related functions or related ligands. Comparative analyses were also carried out on three sets of proteins with diverse functions and ligands. The following describes the details of these sets, then the computational similarity methods, and finally the statistical analysis approaches.

### 2.3.1 Molecular Data Sets

#### 2.3.1.1 Related Proteins

Two sets of kinases were used in this study. One was the set of 45 protein-ligand structures curated in the BindingMOAD database[17] that corresponded to EC number 2.7.11.24, of which 39 were human proteins, 4 mouse, and 2 rat. The second kinase set was obtained from a previous work on binding site comparisons by Kinnings and Jackson.[7] Their study analyzed 351 structures of 76 different kinases, of which 316 structures spanning 64 different kinases had bound ligands, and of those we employed 183 structures representing 26 human kinases for which binding affinities were available. We focused on the set of structures containing bound ligands in order to simplify specification of binding site location. Structures for the ATP data set were obtained from the RCSB PDB database.[18] The structures were obtained from a query for all structures, returning those structures containing a ligand identified as ATP and passing the PDB sequence similarity filter of 95%. This resulted in 267 protein crystal structures (36 archaeal, 125 bacterial, 94 eukaryotic, 10 viral, and 2 synthetic), which were inspected manually to ensure that in every structure ATP was inside a binding site. Of note, the set was dominated by proteins with very low sequence similarity, with 95% of all protein pairs from the set having less than 20% sequence similarity, assessed by global sequence alignment by Needleman-Wunsch.

#### 2.3.1.2 Diverse Proteins

Three sets of heterogeneous protein structures reported by Kahraman et al.,[15] Hoffman et al.,[8] and Yeturu and Chandra[6] were used to make direct comparisons between PSIM and other methods. The Kahraman Set considered 100 protein structures, comprising multiple subsets of sequence-dissimilar proteins that each bound the same ligand type, as follows: PO4 (20 structures), Heme (16), NAD (15), ATP (14), FAD (10), AMP (9), FMN (6), glucose (5), and sex hormones (5). This set was characterized by significant diversity in both ligand size and in corresponding overall binding pocket volumes. The Hoffman Set was designed specifically to mimic the Kahraman set, but to limit the effects of diverse ligand sizes and pocket volumes on the binding pocket comparisons. It consisted of 100 protein structures, with 10 examples for each of 10 ligands (PDB ligand codes follow): 1PE, BOG, GSH, LDA, LLP, PLM, PMP, SAM, SUC, and U5P. The Yeturu Set included multiple binding sites for each of four ligands (methotrexate, indinavir, citrate, and phosphoglycolic acid), but the alternate binding sites were generally highly similar, including, for example, symmetry-related sites within a single protein structure.

### 2.3.2 Computational Methods

The basic notion behind our approach to molecular comparison is that we ought to make comparisons of molecules based on what their binding partners see. For ligands, we want to compare them based on the surfaces moieties that can be recognized by proteins. Given an alignment of two molecules, we define a similarity function that compares distances to the molecular surfaces from observer points surrounding the molecules. Computing the similarity requires identification of the alignment that maximizes this function. The observers are placed on a uniform grid of points with spacing $\lambda$. The points are weighted based on their minimum distance to the molecular surface, retaining a set of observers that correspond closely to a chosen distance $\gamma$ (sharpness is controlled by $\omega$). This identifies a finite set of observer points that are all outside the molecules. Where molecular surfaces are largely congruent in terms of both shape and polarity, the observer points will see the same things in the optimal alignment between the molecules.

Figure 2.1 illustrates the concept. In the case of small molecule ligands, we use 2.0Å, 4.0Å, and 0.2 for $\lambda$, $\gamma$, and $\omega$, respectively. The similarity function itself is a normalized sum of Gaussian functions of the differences in distance from each observer point to each molecule's surface. Such differences are computed for the minimum distance to any surface point (which gives the molecular shape), the minimum distance to a donor surface or formally positive atomic surface, and the minimum distance to an acceptor or negatively charged surface. Directionality and charge magnitude are also taken into account. Details can be found in the original article.[10] The overall effect is that following alignment optimization, for molecules that can exhibit very similar molecular surfaces, both in terms of shape and disposition of polarity, the similarity function will return a value close to 1 whether or not the underlying molecular scaffolding is similar. The function itself is continuous and piecewise differentiable, which makes it suitable for computational optimization. For small molecules, where both conformational flexibility and relative alignment must be optimized, the proce-

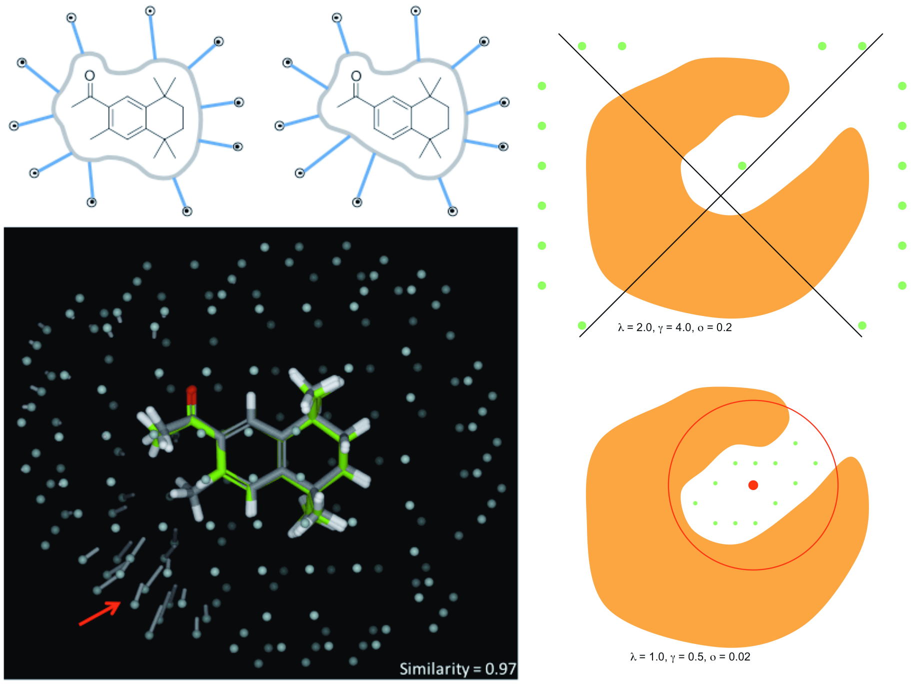


Figure 2.1: Molecular shapes can be characterized by the distances to the molecular surface from points in space. The differences in these distances form the basis for comparison between molecules. At top left, two molecules are cartooned with distances from observers placed outside their surfaces. The differences between the molecules are depicted, lower right (red arrow), with rods of specific lengths corresponding to the differences in distances from observers. Using a normalized Gaussian function of the distance differences, a similarity function is defined whose optimum rewards surface concordance. At top right, the limitation of the approach for comparing protein binding sites is shown. With the observer parameters set for ligands, the binding site is not characterized. By changing the parameters and specifying a radial scope, the binding pocket is densely sampled by observer points.

dure involves a divide-and-conquer strategy to address the conformational problem, heuristic search approaches to address the gross alignment, and local gradient-based optimization for final pose refinement.[10,11]

As seen in Figure 2.1, the choices for $\lambda$, $\gamma$, and $\omega$ yield sensible results for ligands, where we wish to compare the outside of one ligand with the outside of another. For proteins, these values do not typically lead to sensible characterization of a specific pocket. At right in Figure 2.1, the mapping to protein site comparison is shown. By choosing a tighter grid ($\lambda = 0.5$), at a closer spacing ($\gamma = 0.5$), with a thinner shell of highly weighted points ($\omega = 0.02$), within a specified radius of a point within the pocket in question, we can achieve the desired behavior: local comparison of concavities. For the experiments in this article, the point used is the centroid of a co-crystallized ligand, but ligands are not required and automated pocket detection could just as easily designate an approximate pocket center. For protein similarity computations, conformational variation is not explored, and the alignment optimization is carried out through sampling orientations quite densely to identify reasonable starting points for gradient-based local optimization of the similarity function.

Figure 2.2 shows how the approach is applied to two divergent human kinases (CDK2 and c-MET), which share less than 20% sequence identity but nonetheless have common ligands such as staurosporine. These proteins were brought into alignment by maximizing our local pocket similarity metric. Local sequence changes (proline and tyrosine replaced with glutamine and phenylalanine) yielded relatively little difference in the surfaces presented by these pockets. In Figure 2.3, a visualization of the similarity comparison is presented. The left image shows the placement of the observation points from the pocket alignment (superimposed around the transformed ligands) while the image on the right visualizes the similarities observers from these points. Red sticks represent similar negatively charged surfaces, blue represent similarities in positive surfaces, and green sticks correspond to similarities in hydrophobic surfaces. The kinase hinge region responsible for binding ATP contains the most congruent parts of both pockets.

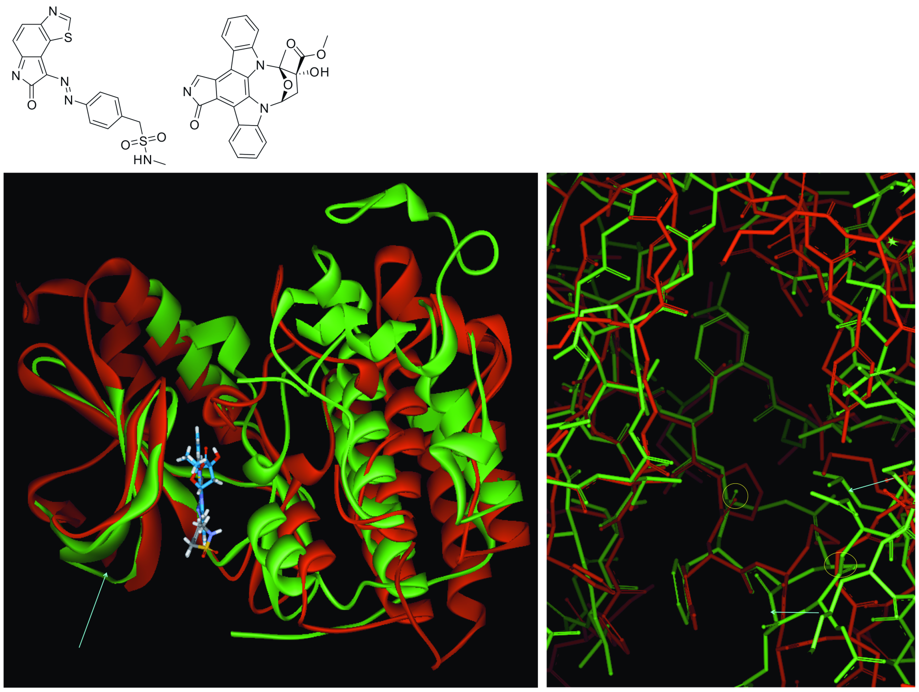


Figure 2.2: Two proteins are shown: CDK2 (1KE6, green, top left ligand) and c-Met (1R0P, red, top right ligand). They have modest sequence identity (less than 20%) and significant differences in overall structure at a global scale, especially in the right-hand lobe (evident at left). However, their binding sites are quite similar in structure (enlarged at right), enough so that they both bind staurosporine. In the hinge region, c-Met makes use of a proline and tyrosine and CDK2 makes use of a glutamine and phenylalanine (blue arrows), but the surfaces are similar enough that both enzymes will bind staurosporine analogs in similar orientations, with analogous hinge binding interactions (hinge acceptor and donor are circled in yellow).

### 2.3.3 Computational Procedures

Detailed scripts for generating the results presented here are available in the data archive associated with this article. Briefly, the procedures included automatic conversion of primary PDB files into mol2 format in order to address bond-order and protonation for both proteins and extracted ligands. All-by-all comparisons of proteins with pocket locations identified by ligand binding sites were performed automatically using the procedure outlined above. Such comparisons yielded similarity scores and alignment transforms among each pair of protein pockets. These were used to build

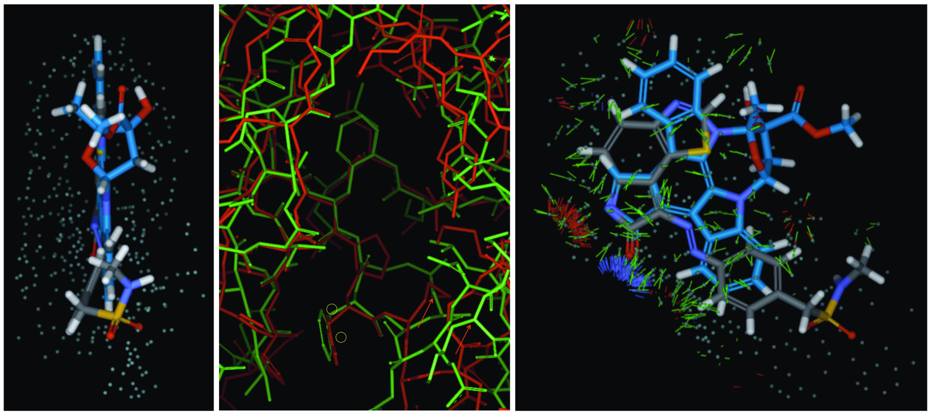


Figure 2.3: The protein alignment of c-Met to CDK2 was computed from the observers shown here outside the ligands (left panel). The right panel shows the relative alignment of the ligands (viewed from the left side of the left and middle panels). The analogous polar interactions of the two ligands to the hinge region of the kinases (yellow circles, red arrows) manifest as an area of high similarity between the proteins. The overall binding pocket shapes are also relatively concordant (green sticks). The cognate ligand of the c-Met structure was closely related to staurosporine (blue carbons), which itself is a potent CDK2 inhibitor. The relatively high similarity in active sites between c-Met and CDK2 is exhibited both directly in the surfaces of their ATP binding sites as well as in the ligands that bind them.

fully aligned trees of protein structures using a greedy approach, adding proteins to a growing tree by seeking the next highest similarity for a protein outside the tree, using the proper pairwise transform combinations to bring all proteins into mutual alignment. Care was taken to define binding site locations and scope using uniform methods to allow for automatic induction of protein alignment trees from initially unaligned proteins.

In the comparison of proteins of closely related function, we employed parameters for the similarity computation of $\lambda = 0.5$, $\gamma = 0.5$, and $\omega = 0.02$, with even weighting of hydrophobic and polar features. We made use of a single definition of binding site scope from the final joint mutual alignment of all proteins. This was done for both kinase sets and for the ATP set. For comparing pairs of proteins from sets of widely divergent character, we employed parameters for the similarity computation of $\lambda = 0.5$, $\gamma = 1.0$, $\omega = 0.02$, and a 0.5 weighting of polar features relative to hydrophobic features. The binding site scope was the union of scopes from each protein site within a given pair. Binding site scope for each protein site was focused on the interaction zones between ligand and protein. This definition resulted in only a weak relationship to pocket volumes, focusing instead on local chemical surface characteristics.

In what follows, Surflex-PSIM will be abbreviated as PSIM for the sake of brevity.

## 2.4 Results and Discussion

We considered three data sets from closely related proteins and three sets from heterogeneous proteins and will discuss results for the two classes in sequence. Among related proteins, we applied the PSIM computation to three different levels of protein structural diversity. The most closely related protein structure set was derived from BindingMOAD, containing all protein-ligand complexes with EC number 2.7.11.24

(mitogen activated protein kinases). This comprised 45 structures of three different kinases (p38$\alpha$, JNK3, and ERK2). We then considered a larger and more diverse set of human protein kinases based on the work of Kinnings and Jackson,[7] composed of 316 structures of 64 different protein kinases. Last, we considered a set of 267 protein structures, all bound to ATP, but spanning highly divergent protein families (e.g. ATP-binding domains of ABC transporters and DNA/RNA polymerases).

### 2.4.1 MAP Kinases

The set of 45 MAP kinase structures included three gene products: 29 MAPK14 (p38$\alpha$), 8 MAPK1 (ERK2), and 8 MAPK10 (JNK3). We computed an all-by-all comparison of these structures using PSIM, and Figure 2.4 shows the resulting tree of similarity. The different kinase families were segregated well, with distinctions also made between different species and mutant proteins. The most similar pair (1WBT and 1WBS) contained wild-type human p38$\alpha$ bound to nearly identical ligands, differing only by a carbon/nitrogen swap in a heterocycle.

Following the nodes away from the root of the tree, we continue to see a substructure of the first two ligands bound to 1WBV, with excursions from the 1WBT ligand envelope occurring as we move up the tree. The ligand of 1OVE makes two separate protrusions from the binding envelope of the 1WBT ligand. Pocket conformations of 1WBS, 1WBV, and 1OVE are shown relative to 1WBT at left in Figure 2.4. The very different (and rigid) ligand of 1OVE binds to a different DFG configuration of the kinase altogether (the inactive conformational form). However, the structure is still correctly grouped with the p38 exemplars, and it is correctly aligned with respect to the common hinge-binding elements shared by all of the ligands.

To formalize the relationship between the taxonomy represented by the tree in Figure 2.4 and ligand structures, we computed similarities for different groups of p38 ligands based on the tree structure. The set of protein conformational variants rooted at 1WBT and including those proteins up to depth 4 (seven structures total) defined

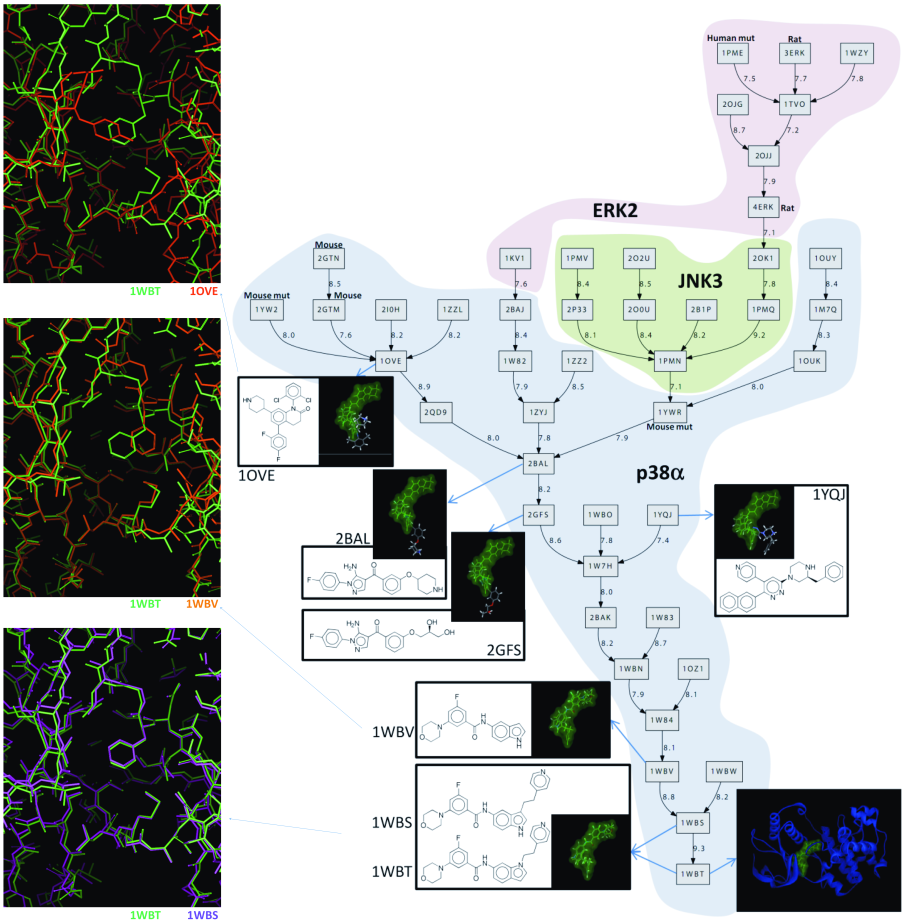


Figure 2.4: The alignments of protein pockets from EC family 2.7.11.24 are shown in a single-linkage hierarchical clustering. Values along links indicate pocket similarity. The ligands all bind in the hinge region of the kinases. The three different kinases are nearly perfectly segregated based on the pocket similarity alone. Among the p38$\alpha$ variants, pocket similarity agrees in a qualitative sense with ligand similarity. The 1WBT ligand and corresponding surface are shown in all snapshots. Three different pocket conformations are shown superimposed on to 1WBT at the bottom, illustrating the increasing conformational change as one moves further from the root of the tree.

a group of ligands that were bound to highly similar pockets. The set of p38$\alpha$ variants at depth 10 or greater (12 structures, at the same level as or above 1OVE in Fig. 4) defined a group of ligands bound to dissimilar pockets from those near 1WBT. Pairwise three-dimensional ligand similarities among the pocket group near 1WBT were higher than those between that group and those distant from 1WBT (ROC area 0.77, $P \ll 0.01$ by resampling). Considering the ranked list of ligand similarities, pairs at the top of the list were over 20-fold more likely to come from proteins belonging to the group near 1WBT than to the cross-pairs that included a ligand from near the root and one distant from it. This will be further quantified below in comparison with other methods.

The resulting alignments have two desirable properties. First, similarity between protein pocket variants of the same flexible enzyme is related to the similarity of ligands that bind the pocket variants. Second, alignments that optimize local surface similarity preserve the geometry of parts of the protein surface that remain congruent when other parts of the protein binding pockets change significantly. The second property stems from the definition of the similarity metric as one that rewards similarity as opposed to penalizing differences.

### 2.4.2 Kinase Ligand Binding Profiles: Kinnings Data Set

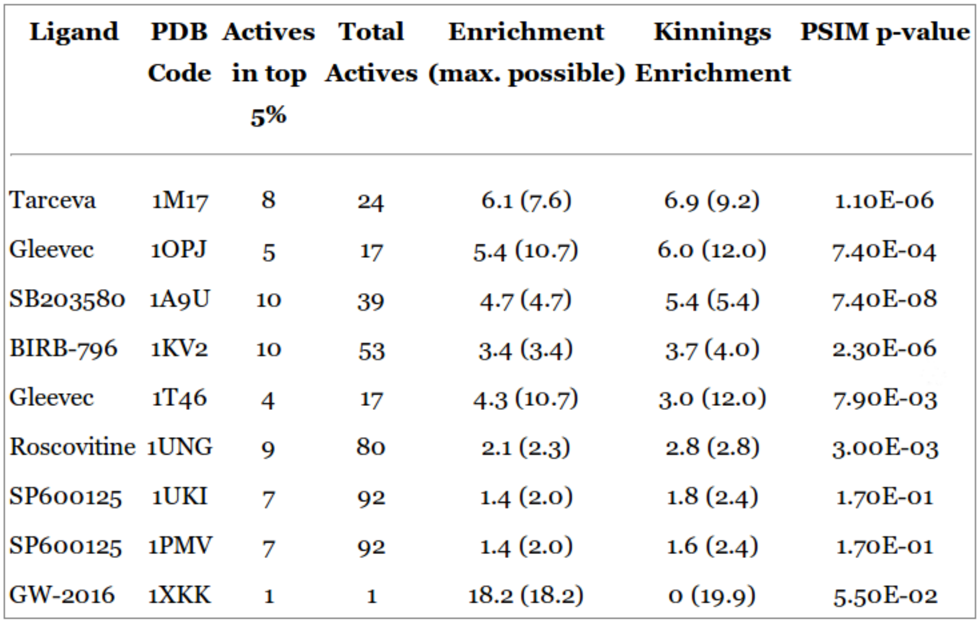
One of the main goals of the structural analysis of proteins is the ability to differentiate proteins based upon their ligand binding preferences. The 183 structure/26 protein human kinase data set presents a particularly appealing target for differentiating ligand binding because of the difficulty of designing inhibitors with high specificity and the value of designing inhibitors with such specificity for cancer research. The work done in Kinnings and Jackson[7] utilized enrichment testing for judging the success of their similarity metric as a methodology for determining the binding preference of proteins.

Given a query ligand Y, the idea was to identify proteins that would also bind Y by comparing their structures with the structure of a protein bound to Y. Enrichment was calculated based on the number of structures in the top 5% of the ranked list whose corresponding protein was inhibited the target ligand with a Ki of less than 10 $\mu$M. The enrichment score was defined as ((Ah/Th)/(A/T)): Ah was the number of structures which bind the target ligand in the top 5%; Th, the total number of structures in the top 5%; A, the total number structures which bind the target ligand; and T, the total number of structures. P values for enrichment scores were computed using a hypergeometric distribution. Computing enrichment in this fashion, PSIM yielded nearly identical results to those of Kinnings (Table 2.1).

However, it is important to understand that the results for a number of cases are dominated by the presence of multiple alternative protein structures for the cognate protein of the query ligand. The test considered all protein structures as being separate individuals, so even those structures that represented alternate conformations of the same protein were considered as being distinct in the enrichment testing. So, given a structure Z of ligand Y bound to protein X, an alternate conformation of protein X corresponding to structure Z would positively influence the enrichment associated with ligand Y. Enrichment scores computed in this fashion can be somewhat misleading. For example the target structure for the ligand SB203580 is 1A9U, which is a crystal structure of p38$\alpha$ bound to SB203580. The top 5% of structures similar to 1A9U are also crystal structures of p38$\alpha$, entirely dominating the enrichment computation (p38$\alpha$ has 34 representative structures in the data set).

This bias is not apparent in all of the ligands tested. For example, results for Tarceva showed a significant enrichment factor that was obtained by the high ranking of structures from several different proteins. The target structure for the enrichment

Table 2.1: Comparison of PSIM to results from Kinnings et al. The table shows enrichment for structures whose corresponding enzyme was known to bind the cognate ligand of the query PDB structure. The results were nearly identical (minor differences in the numbers of proteins were due to technical differences in the protocols). Note, however, that enrichment values (e.g. for SB203580, BIRB-796, and Roscovitine) are dominated by multiple structural variants of the cognate enzyme being present in the analysis (see text for details). For Tarceva, by contrast, the highly ranked structures included some with low sequence similarity to EGFR (the protein in 1M17).



| Ligand | PDB Code | Actives in top 5% | Total Actives | Enrichment (max. possible) | Kinnings Enrichment | PSIM p-value |
|---|---|---|---|---|---|---|
| Tarceva | 1M17 | 8 | 24 | 6.1 (7.6) | 6.9 (9.2) | 1.10E-06 |
| Gleevec | 1OPJ | 5 | 17 | 5.4 (10.7) | 6.0 (12.0) | 7.40E-04 |
| SB203580 | 1A9U | 10 | 39 | 4.7 (4.7) | 5.4 (5.4) | 7.40E-08 |
| BIRB-796 | 1KV2 | 10 | 53 | 3.4 (3.4) | 3.7 (4.0) | 2.30E-06 |
| Gleevec | 1T46 | 4 | 17 | 4.3 (10.7) | 3.0 (12.0) | 7.90E-03 |
| Roscovitine | 1UNG | 9 | 80 | 2.1 (2.3) | 2.8 (2.8) | 3.00E-03 |
| SP600125 | 1UKI | 7 | 92 | 1.4 (2.0) | 1.8 (2.4) | 1.70E-01 |
| SP600125 | 1PMV | 7 | 92 | 1.4 (2.0) | 1.6 (2.4) | 1.70E-01 |
| GW-2016 | 1XKK | 1 | 1 | 18.2 (18.2) | 0 (19.9) | 5.50E-02 |

of Tarceva was 1M17 (EGFR bound to Tarceva). The top 5% most similar structures contained structures of the proteins SRC and LCK. LCK has one of the most similar sequences to EGFR of the proteins in the set, but it was ranked below SRC, which has much less sequence similarity.

We made an alternative analysis that removed the bias of multiple structural exemplars by defining the similarity of two proteins as the maximum similarity obtained using the PSIM method computed over all pairs of structures of the two proteins. We employed receiver operator characteristics (ROC) analysis to determine whether pairs of proteins which both bind the same ligand had significantly higher similarity than protein pairs where one binds a particular ligand and the other does not. Resampling was used to determine the significance of the ROC areas. In 6/10 cases, the ROC area exhibited significant separation ($p < 0.05$), showing higher similarity among protein pairs that shared ligands. The ROC plot for Tarceva is shown in Figure 2.5 (ROC area 0.77, $P < 0.002$). The alignment of EGFR with SRC gives a sense of the high local surface similarity that drives the high score (green arrows) along with significant differences that reflect the relatively low sequence similarity of the two proteins (red arrows).

The combination of sequence divergence coupled with surface epitope preservation is a common theme in biology. As sequence divergence increases, such three-dimensional structural motifs are detectable with a similarity metric that focuses on surfaces and ignores sequence information.

### 2.4.3 Motif Discovery: ATP-Binding Data Set

The detection of motifs in proteins was originally and is still primarily done through the identification of sequence-based motifs or the computation of sequence similarity to known protein domains. These motifs are then used for the annotation and classification of unknown proteins. While these methods are extremely effective for long sequence motifs and proteins that share significant evolutionary history, they have limited ability to detect short discontinuous motifs and protein similarities based on convergent structures rather than amino acid homology from shared ancestry. Utilizing a three-dimensional structure-based method allows for the discovery of motifs that are discontinuous and may reduce false attribution when sequences are similar but a small change has significantly altered the structure.

ATP is an attractive ligand for considering distantly related proteins because of the long history of the molecule with respect to the evolution of life. Protein motifs

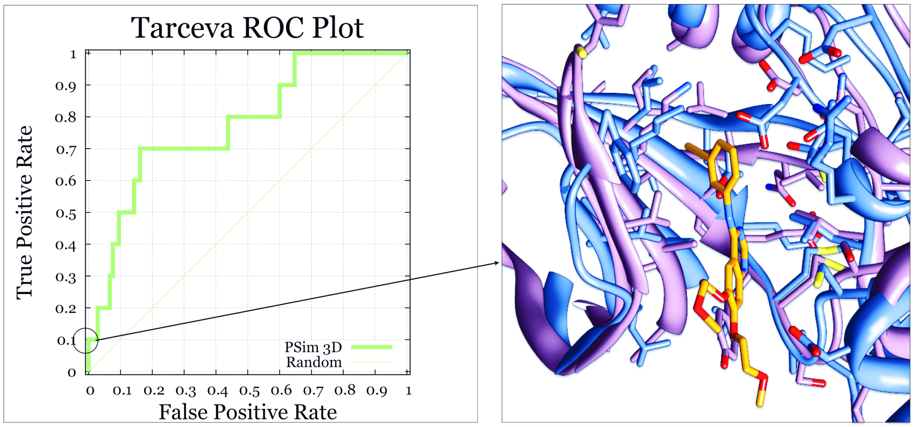


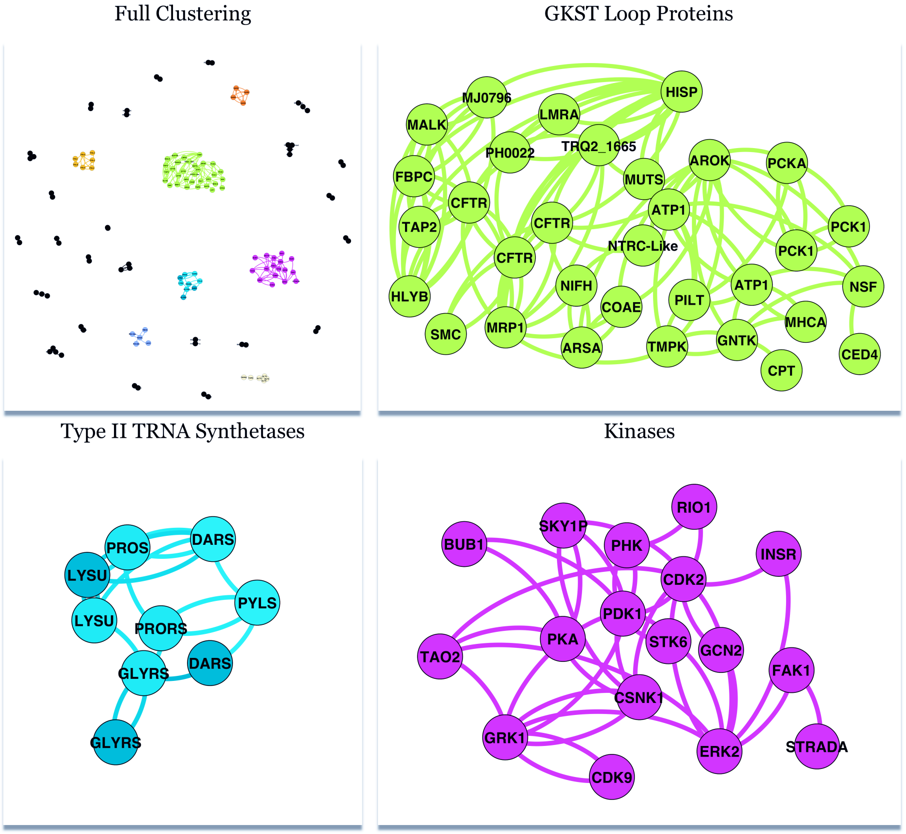
Figure 2.5: By defining protein similarity as the maximum similarity over all pairs of conformational variants between two proteins, one can directly measure enrichment for high pocket similarities among proteins that share ligands. Here, similarities of pairs of proteins known to be inhibited by Tarceva are compared with similarities between such proteins and proteins not inhibited by Tarceva. The ROC plot shows significant separation, indicating higher similarity among the protein pairs sharing Tarceva compared with pairs that do not. At right, the most similar pocket variants of EGFR and SRC (the single highest protein similarity among those computed for Tarceva) are shown in their optimal alignment (EGFR in blue and SRC in purple).

that bind ATP have been projected to be among the very first to have evolved.[19] This timeframe allowed for the development of many parallel ways of utilizing the molecule and also significant time for convergent pathways to produce congruent structures. Such common structural motifs can be discovered using a method analogous to the discovery of protein families with sequence similarity. Instead of grouping proteins together based upon common sequence and extracting a motif, proteins can be clustered based upon common binding site structure to reveal common structural motifs.

The 267 ATP-bound structures from the PDB were a particularly diverse set, including 36 archaeal, 125 bacterial, 94 eukaryotic, 10 viral, and 2 synthetic proteins, with 95% of the pairwise sequence similarities being less than 20%. Using PSIM, clusters were created from the ATP data set by first forming a fully connected graph with protein structures as nodes and PSIM similarity values as the weights on edges. The edges were then filtered to only retain highly significant ($p < 0.001$) edges based on a similarity threshold derived from a set of  100,000 randomly selected unrelated protein pair similarities. The resulting clusters where then annotated based upon known protein functions. Figure 2.6 depicts the overall cluster and highlights subclusters of particular interest.

#### 2.4.3.1 Kinases

One of the prominent clusters contained many of the structures from the work discussed previously. Kinases have been well annotated both for functional sequences and structural motifs. Sequence motifs have been defined for various regions of the binding domain and have previously been used to classify unknown proteins as kinases. Without the aid of this prior information, based purely on local binding site surface similarity, PSIM separated the kinases as a distinct cluster (Fig. 2.6, lower right). The cluster also included the Pseudo-Kinase STRAD$\alpha$, which contains a known kinase subdomain but lacks some required catalytic residues, suggesting that the methodology is clustering based on binding modality and not catalytic activity. When the kinases are put into a common alignment induced from their surface similarity, a pronounced pocket is visible (see Fig. 2.7) showing the consensus structure used by these kinases to bind ATP. The most conserved portions of this surface are


Full Clustering

GKST Loop Proteins

Type II TRNA Synthetases

Kinases

Figure 2.6: The full clustering of the diverse ATP-bound protein set (upper left), yielded two large clusters: GKST loop proteins (upper right) and kinases (lower right). A number of small clusters were also identified, of which the tRNA Synthetases (lower left) were typical. The connections in the clusters represent those edges of the fully connected graph among all of the ATP-bound proteins whose significance exceeded an estimated P value of 0.001.

located in the nucleotide-binding region while greater variability exists near the phosphate tail. The similarity in all of the kinase binding sites stems from the common binding mode used by these proteins in their interactions with ATP.

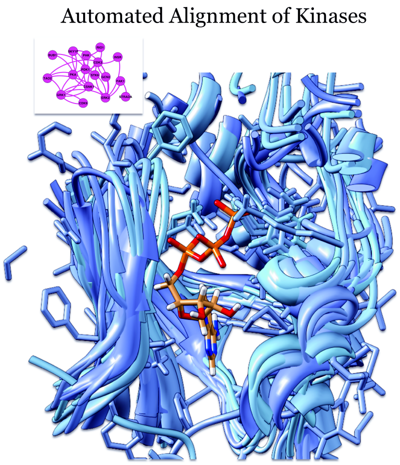


Figure 2.7: The alignment of all of the structures in the kinase cluster is shown with a single bound ATP to indicate the binding site. Variability increases among the proteins near the phosphate tail but the area around the nucleotide head corresponding to the hinge-binding region is strongly conserved.

However, different protein families make use of different tricks in binding ATP. Figure 8 highlights the binding modes of ATP within three different clusters from the global clustering. For kinases (lower right), disposition of the nucleotide head is largely conserved, mirroring similarities in binding the hinge region of kinases (so-called since it serves as a hinge between two protein domains, as seen in Fig. 2.2). The phosphate tail exhibits greater variability, apparently reflecting the differing specificity of kinases with respect to transfer of phosphate to substrates. By contrast, the GKST loop proteins are strongly conserved in the phosphate tail region with great variability in the nucleotide portion, and the tRNA synthetases show a more closely conserved ATP binding mode overall that is different from both. Note that since PSIM rewards surface congruence (as opposed to minimizing overall deviations), those regions of the binding sites that are most similar can be recognized without being disrupted by the more variable regions.

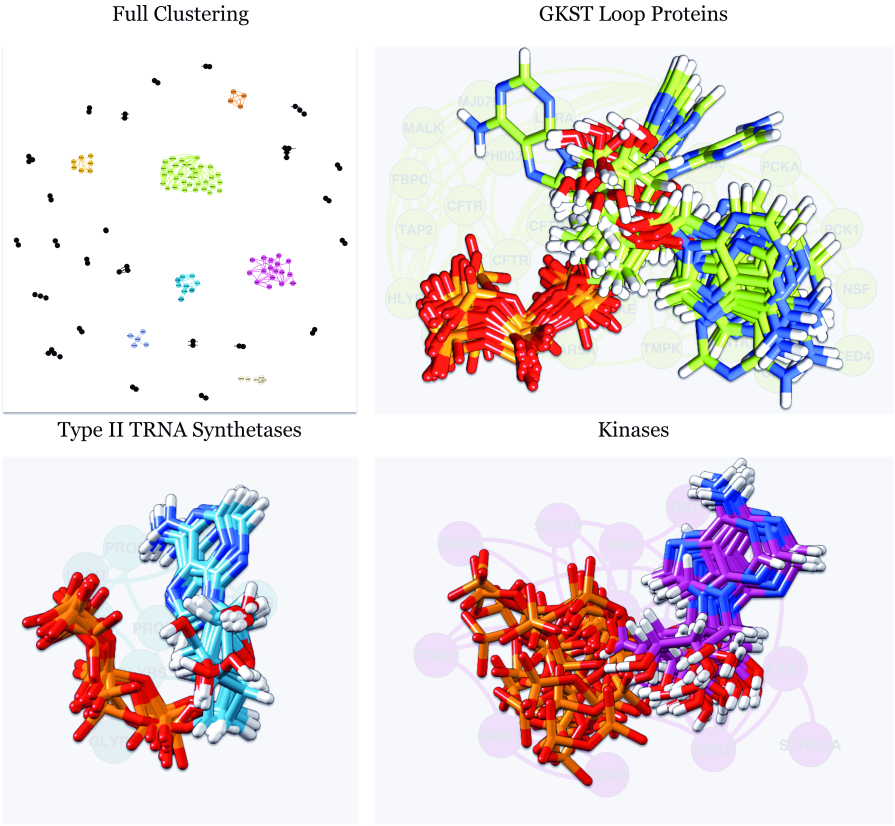


Figure 2.8: The conformational variations of ATP in the alignments derived from PSIM computations are shown for the three clusters highlighted in Figure 2.6. The GKST group appears focused on consistent binding of the phosphate tail of ATP, the Kinases are focused on the binding the nucleotide portion, and the tRNA Synthetases bind both portions in a consistent way that is divergent from both of the previous groups.

Although many of the known sequence features for kinases are short and discontinuous,[20] a proper structural alignment should overlay the common residues making sequence based motif discovery more tractable. Such an alignment also guarantees that the residues not only share similar placement in sequence space but also share a similar relative structural location and enhances the chance that function is also shared. Extraction of sequence-based motifs from sequence alignments derived from the PSIM aligned structures produced many well-known features of kinases. Sequence-based motifs were analyzed using UCSF Chimera[21] on the PSIM kinase pocket alignment. Motifs representing the DFG activation loop and the invariant lysine implicated in catalytic rate were found as well as variations on the well-known hinge-binding motifs[22] (see Fig. 2.9). These small motifs would be nearly undetectable in a sequence based approach applied to as diverse a group of proteins as our ATP set. Motifs such as GXGXXG would have been particularly difficult due to the lack of specificity but a structure shaped by this motif in these protein contexts provides ample specificity to be discovered with the assistance of a three-dimensional method.

#### 2.4.3.2 GKST Loop

One of the oldest motifs known (both historically and evolutionarily) is the P-loop-containing triphosphate hydrolase fold,[19] which has been noted for its conserved GKST motif. The loop is also known as the Walker-A motif. This short and weakly defined sequence motif (GXXXXGK[TS]) would be difficult to extract from a set of proteins as diverse as the ATP set based on sequences alone. We observed, however, that the corresponding structures yield an extremely well defined binding motif for ATP. The largest cluster (Figs. 2.6 and 2.8, green cluster) that PSIM generated corresponded to this motif and exhibited a remarkably closely conserved tail conformation for ATP (see Fig. 2.8). The PSIM clustering correctly grouped proteins with widely varying evolutionary history. Figure 2.10 shows how the GKST structural loop binds the phosphate tail of ATP in conjunction with a magnesium ion. All of the phosphate tails exhibit a nearly identical orientation with the nucleotide head free to be handled differently depending on the protein involved.

The remarkable structural conservation comes despite very significant sequence variation. Figure 2.11 shows the difference between the sequence alignment derived from the PSIM structural alignment and that derived by Clustal W[21] purely based on sequence. GKST stands out as a significantly conserved motif but we also gain

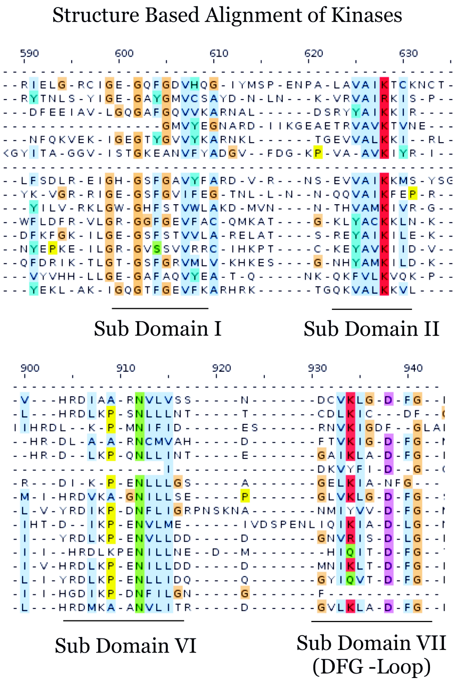


Figure 2.9: The sequences of the kinase cluster proteins aligned based upon their structures. Subdomain I has been noted act as a clamp that anchors the non transferable phosphates of ATP. Subdomain III contains an invariant Lysine that interacts with the alpha and beta phosphates and is required for maximal enzyme activity. Subdomain VI has strongly conserved residues but has not been annotated as being related to catalysis or substrate interactions. Subdomain VII contains the DFG loop which is used in kinase activation.

insight into some of the more drastic changes that can occur within this motif. In some cases, Clustal W was able to correctly identify the sequence motif, but in many cases, the motif was too weak. This is most likely due to the shortness of the GKST motif and the presence of many other similar regions in the proteins in question. The example of PDB structure 3FKQ from E. Rectale is particularly striking, since it is the only protein lacking the highly conserved lysine within the entire set. In this case the lysine was replaced with a threonine but ATP still binds, and does so in a manner nearly identical to those proteins bearing the canonical motif. Clustal W yields an unrelated sequence for 3FKQ in the multiple alignment, but the pocket similarity is sufficiently strong (p <0.001) that PSIM was able to both place the protein in the correct cluster as well as identify the correct alignment correspondence with the other proteins in the family.

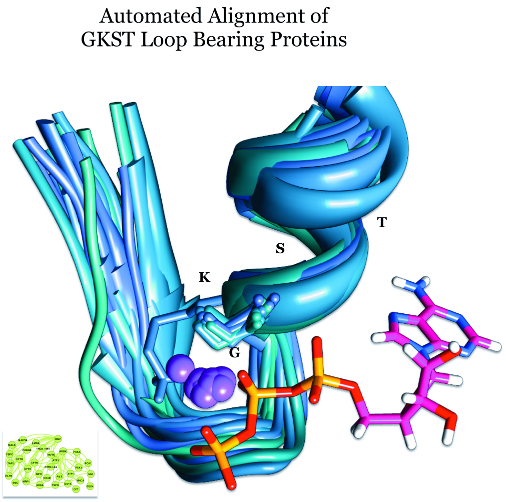
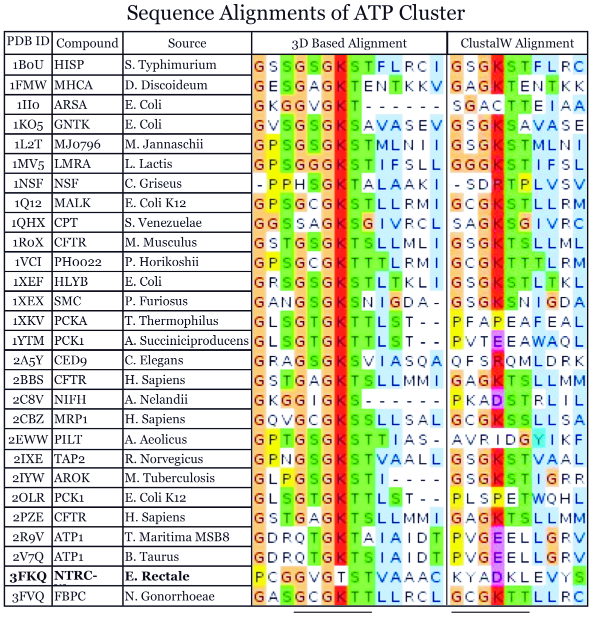


Figure 2.10: The PSIM alignment of the proteins from the GKST loop cluster of Figure 2.6 are shown. Despite a great deal of local sequence variation and significant global sequence and structural variation, the loops form a very consistent structure, making use of a magnesium ion (shown in purple) to bind the phosphate tail of ATP.

### 2.4.4 Heterogeneous Protein Data Sets

Among heterogeneous proteins, we applied PSIM to three different data sets (see Methods and Data for details), each constructed by different research groups reporting new protein pocket similarity approaches, and with each addressing slightly different considerations. The Kahraman Set considered 100 protein structures, comprising multiple subsets of sequence-dissimilar proteins that each bound the same ligand



Sequence Alignments of ATP Cluster

| PDB ID | Compound | Source | 3D Based Alignment | ClustalW Alignment |
|---|---|---|---|---|
| 1BoU | HISP | S. Typhimurium | GSSGSGKSTFLRCI | GSGKSTFLRC |
| 1FMW | MHCA | D. Discoideum | GESGAGKTENTKKV | GAGKTENTKK |
| 1IIo | ARSA | E. Coli | GKGGVGKT------ | SGACTTEIAA |
| 1KO5 | GNTK | E. Coli | GVSGSGKSAVASEV | GSGKSAVASE |
| 1L2T | MJ0796 | M. Jannaschii | GPSGSGKSTMLNII | GSGKSTMLNI |
| 1MV5 | LMRA | L. Lactis | GPSGGGKSTIFSLL | GGGKSTIFSL |
| 1NSF | NSF | C. Griseus | -PPHSGKTALAAKI | -SDRTPLVSV |
| 1Q12 | MALK | E. Coli K12 | GPSGCGKSTLLRMI | GCGKSTLLRM |
| 1QHX | CPT | S. Venezuelae | GGSSAGKSGIVRCL | SAGKSGIVRC |
| 1RoX | CFTR | M. Musculus | GSTGSGKTSLLMLI | GSGKTSLLML |
| 1VCI | PH0022 | P. Horikoshii | GPSGCGKTTTLRMI | GCGKTTTLRM |
| 1XEF | HLYB | E. Coli | GRSGSGKSTLTKLI | GSGKSTLTKL |
| 1XEX | SMC | P. Furiosus | GANGSGKSNIGDA- | GSGKSNIGDA |
| 1XKV | PCKA | T. Thermophilus | GLSGTGKTTLST-- | PFAPEAFEAL |
| 1YTM | PCK1 | A. Succiniciproducens | GLSGTGKTTLST-- | PVTEEAWAQL |
| 2A5Y | CED9 | C. Elegans | GRAGSGKSVIASQA | QFSRQMLDRK |
| 2BBS | CFTR | H. Sapiens | GSTGAGKTSLLMMI | GAGKTSLLMM |
| 2C8V | NIFH | A. Nelandii | GKGGIGKS------ | PKADSTRLIL |
| 2CBZ | MRP1 | H. Sapiens | GQVGCGKSSLLSAL | GCGKSSLLSA |
| 2EWW | PILT | A. Aeolicus | GPTGSGKSTTIAS- | AVRIDGYIKF |
| 2IXE | TAP2 | R. Norvegicus | GPNGSGKSTVAALL | GSGKSTVAAL |
| 2IYW | AROK | M. Tuberculosis | GLPGSGKSTI---- | GSGKSTIGRR |
| 2OLR | PCK1 | E. Coli K12 | GLSGTGKTTLST-- | PLSPETWQHL |
| 2PZE | CFTR | H. Sapiens | GSTGAGKTSLLMMI | GAGKTSLLMM |
| 2R9V | ATP1 | T. Maritima MSB8 | GDRQTGKTAIAIDT | PVGEELLGRV |
| 2V7Q | ATP1 | B. Taurus | GDRQTGKTSIAIDT | PVGEELLGRV |
| **3FKQ** | **NTRC-** | **E. Rectale** | PCGGVGTSTVAAAC | KYADKLEVYS |
| 3FVQ | FBPC | N. Gonorrhoeae | GASGCGKTTLLRCL | GCGKTTLLRC |

Figure 2.11: Sequence alignments are shown for those derived from the PSIM structural alignment (left) and using ClustalW for pure sequence based alignment (right). ClustalW used sequences from the same protein chains used to do the PSIM alignment. All of the PSIM alignments correctly aligned the canonical GK(S/T) central motif (and the single outlier lacking the lysine was still correctly aligned). The sequence-based alignment was clearly incorrect in 11 cases. While it is not surprising that a structure-based method should be more accurate than a sequence-based one, the degree of improvement is striking.

(e.g. ATP, AMP, PO4, and NAD), with significant diversity present in both ligand size and binding pocket volume. The Hoffman Set was designed specifically to mimic the Kahraman Set, but to limit the effects of diverse ligand sizes and pocket volumes on the binding pocket comparisons. The Yeturu Set included multiple binding sites for each of four ligands, but the alternate binding sites were highly similar, including, for example, symmetry-related sites within a single protein structure (e.g. PDB code 1EGH, with six symmetric phosphoglycolic acid binding sites).

#### 2.4.4.1 Kahraman Set

Results from application of a spherical harmonic shape method (the cleft interact version) due to Kahraman et al.[15] and from an atom-cloud comparison method (sup-CK) due to Hoffman et al.[8] measured the ability of a method to rank pockets that bound a particular ligand as being similar to a protein pocket that also bound that ligand. Results were reported using standard ROC area analysis, with the average of areas resulting from 100 different rankings (one each for each protein in the data set). The spherical harmonic approach yielded a mean ROC area of 0.77 (no standard deviation was given).[15] The sup-CK method,[8] using parameters tuned for the Kahraman Set, yielded 0.86 ± 0.14. The PSIM approach yielded 0.79 ± 0.19. None of these methods exhibited differentiable performance from one another, and none performed better than volume alone, which yielded 0.88 ± 0.14.[8] For these proteins, average pocket volume was strongly correlated with ligand size, with PO4 having the smallest binding sites (445 ±118 Å$^3$) and FAD having the largest (2099 ± 224 Å$^3$).[15] Of the three methods, the PSIM approach was least strongly related to volume, with site comparisons being focused on observations of local chemical surface similarities and the binding pocket scope being limited in this computation to the portion of the binding pockets proximal to their cognate ligands.

#### 2.4.4.2 Hoffman HD Set

Given the confusing influence of volume on interpretation of results, the sup-CK authors produced a homogeneous data set: 10 binding sites for each of 10 ligands, but where both the ligand and pocket sizes had far less variance than in the Kahraman Set.[8] On this set, volume alone yielded a mean ROC area of 0.65 ± 0.15, sequence similarity 0.58 ± 0.09, and sup-CK and sup-CKL each yielding tuned performance of 0.71 ± 0.19 and 0.75 ± 0.16, respectively.[8] PSIM yielded 0.76 ± 0.15 using precisely the same parameters as used for the Kahraman Set. While the sup-CK method showed a substantial performance decrease when volume was factored out of the comparisons, the PSIM approach yielded nearly identical performance on both sets.

#### 2.4.4.3 Yeturu Set

Quite a different question was addressed by one of the data sets in the study by Yeturu and Chandra[6] in which the PocketMatch algorithm was introduced. For this set, four different ligands (citrate, indinavir, phosphoglycolic acid, and methotrexate) in 26 total crystal structures (Fig. 3 from the article) yielded 51 binding sites, many of which were multiple minor variants within a single crystal structure. Given that the PSIM approach can make fine distinctions among local pocket differences, sensitivity to natural local variation may have posed some difficulty. However, as shown in Figure 2.12, PSIM showed perfect segregation of the binding sites based on ligand type (using exactly the parameters as used for the Kahraman and Hoffman Sets).

The methotrexate and indinavir cases generally represented very minor variations on nearly identical binding sites. In those cases, similarity scores were nearly all above 9.0. For phosphoglycolic acid, 6 of 11 sites were from a symmetric arrangement within a single structure (1EGH) that yielded pairwise similarities over 9.9. The citrate examples fell into three categories: surface-bound pairs of ligands bound to serine proteases (the bulk of cases), a trio of symmetrically bound citrate ligands bound to a macrophage inhibitory factor protein, and a pair bound to unrelated sites of a

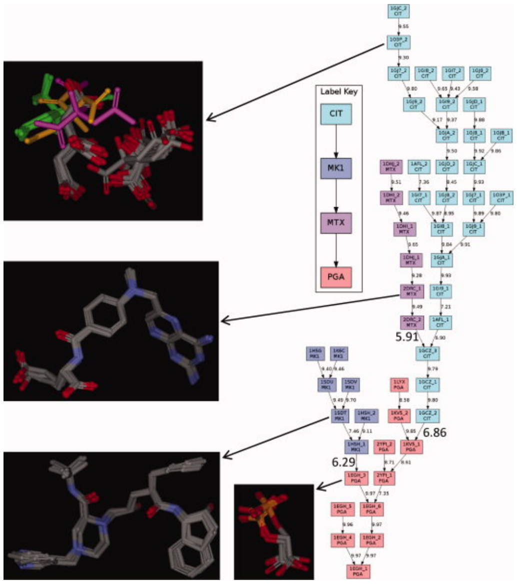


Figure 2.12: The PSIM method perfectly segregated the 51 sites of the Yeturu Set into subtrees of single ligand types. Alignments of the ligands are shown based upon the global alignment tree shown at right. Note that while the PGA (phosphoglycolic acid), MK1 (indinavir), and MTX (methotrexate) sites all represent variations on single binding site themes, the CIT (citrate) sites represent distinct themes. The bulk of the citrate ligands were pairs bound to the surface of serine proteases in a solvent-exposed environment and are shown in atom color. The green citrate alignments come from a trio of symmetric sites within a single protein structure (1GCZ). The orange and magenta alignments come from distinct solvent-exposed citrate molecules from a single ribonuclease structure (1AFL). The edges connecting indinavir to phoshoglycolic acid and methotrexate to citrate had notably low PSIM values, but the edge connecting citrate to phosphoglycolic acid was higher, reflecting genuine similarity between those binding sites (highlighted with larger fonts).

ribonuclease. Despite the heterogeneity of the last set, all segregated into a single subtree. The clustering tree reported by Yeturu and Chandra[6] showed similar results, but grouped the trio of 1GCZ citrate sites within the phosphoglycolic acid subtree and apart from the remainder of the citrate binding sites.

### 2.4.5 Relationship Between PSIM Similarity and Ligand Similarity

A natural expectation is that there should exist some degree of concordance between the similarity of protein pockets and the similarities exhibited by the ligands that bind them. There are two quite different cases within the data sets examined here. One is the case where many synthetic ligands have been designed to competitively bind a single protein's active site (e.g. the 29 different ligands of p38$\alpha$ from Fig. 2.4). The other is the case where nature has evolved multiple strategies for binding the same naturally occurring cognate ligands, which is the situation in the Kahraman Set. Figure 2.13 shows both relationships, with ligand similarity computed using the Surflex-Sim approach.[11]

The plot at left in Figure 2.13 shows ligand similarity versus protein pocket similarity for all pairs of p38$\alpha$ comparisons. The relationship was statistically significant ($p \ll 0.01$ by Kendall's Tau, estimated by permutation analysis), though clearly stronger at the extremes of ligand similarity and dissimilarity. Recalling Figure 2.4, pairs of ligands such as those in 1WBT and 1WBS had very high similarity (9.4) with correspondingly high pocket similarity (9.3), whereas 1WBT compared with 1OVE yielded much lower ligand and protein similarity (4.9 and 6.9, respectively).

The plot at right in Figure 2.13 shows ligand similarity versus protein pocket similarity for the Kahraman data set, where the protein similarity values resulted from average all pairs of comparisons for the cognate proteins of each ligand type and the ligand similarity values were again computed using Surflex-Sim. With the exception of the glucose/phosphate binding site comparison, the relationship between ligand and protein similarity was nearly linear. Overall, the correlation was 0.46 by Kendall's Tau ($P \ll 0.01$, by permutation). Further, the relationship between PSIM values and ligand similarity was much stronger than between PSIM values and volume similarities (Kendall's Tau of 0.30). Even when restricting the set of comparisons to protein pairs where volume differences no longer correlated with protein pocket similarities, the statistically significant relationship between pocket and ligand similarity remained.

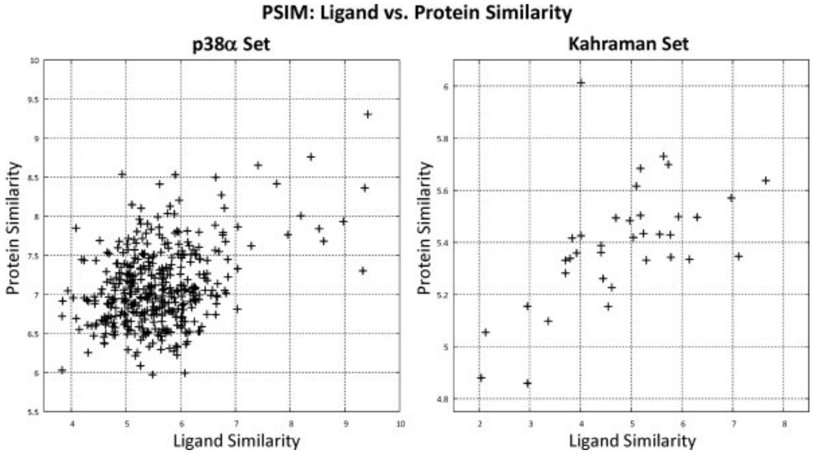


Figure 2.13: The PSIM method exhibits a direct correlation between ligand similarity (x-axis in both plots) and binding pocket similarity (y-axis). For the set of variants of p38 (left), each point represents a single pairwise comparison of protein pockets and of the bound ligands. The pocket differences were driven by differences in the bound ligands (see Figure 2.4, and the correlation was Kendall's Tau = $0.15 (p \ll 0.01)$. For the Kahraman set cross-pairs (right), each point represents the mean pocket similarity of all variants of pairs of different proteins along with the pairwise cognate ligand similarity. Quite diverse families of proteins evolved to bind the natural cognate ligands, but the correlation between pocket and ligand similarity was pronounced ($\tau = 0.46$, $p \ll 0.01$).

The relationship between ligand similarity and pocket configuration within the p38$\alpha$ variants was subtle. To test whether the PSIM approach was unique in its ability to identify this effect, we assessed whether the PocketMatch[6] and SOIPPA[22] methods yielded correlations with ligand similarity for the p38$\alpha$ subset. Figure 2.14 shows the plots of ligand and protein similarity for the two methods. Neither algorithm was able to yield correlations between local pocket similarity and the similarities of bound ligands. When restricted to ligand similarities less than 7.0, PSIM yielded a statistically significant correlation ($\tau = 0.10$, $p < 0.01$), but both PocketMatch and SOIPPA yielded slightly negative correlations (the former with $p < 0.05$).

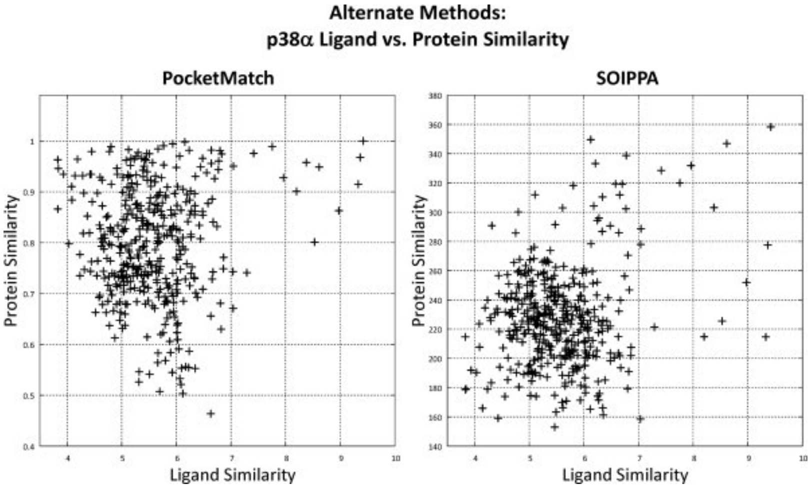


Figure 2.14: The p38$\alpha$ set provides a challenging test of discrimination for subtle pocket conformational effects related to ligand similarity. Neither the PocketMatch algorithm ($\tau = 0.01$, $p = 0.42$), nor the SOIPPA algorithm ($\tau = 0.00$, $p = 0.5$) were able to yield correlations between local pocket similarity and the similarities of bound ligands.

The two cases explored here do not fully examine the issues around the relationship between protein pocket similarity and ligand similarity. The p38$\alpha$ set considered a relatively flexible protein pocket bound to a series of competitive ligands with a range of underlying scaffolding. Clearly, in the case of a very rigid protein, it would be more difficult to discern conformational effects and relate them to ligand structural patterns. There are many examples of proteins that undergo little change on binding structurally divergent ligands. The Kahraman Set was very different in character, with examples of multiple diverse proteins each bound to the same ligand. In this case, the degree of concordance between average pocket similarity and ligand similarity was striking, though the issue of binding site volume differences limits the generality of the observation. It bears mention as well that different approaches to the computation of ligand similarity would also affect the results. Similarity approaches that measure topological structural similarity between ligands, for example, might yield no relationship between pocket and ligand similarity. As with proteins sharing little sequence similarity but exhibiting similar binding pockets, small molecules may share very little topological similarity but exhibit very similar surface shape and polarity.

### 2.4.6 Relationship to Other Approaches

There are a number of protein structural comparison algorithms, broadly characterized by whether they are global or local, backbone-based or include sidechain information, and the degree to which they make comparisons based purely on shape or also based on polarity. The Surflex-PSIM approach is local, accounts for all protein atoms, and considers the detailed comparison of both shape and protein surface polarity. Other methodologies have approached protein similarity in a variety of ways such as geometric hashing, shape alignment, and fingerprinting. These various methods can be local or global and some produce alignments while others do not.

Direct comparisons were made here with five notable recent examples of pocket comparison algorithms: the SiteBase algorithm of Kinnings and Jackson,[7] the spherical harmonics approach of Kahraman et al.,[15] the sup-CK method of Hoffman et al.,[8] the PocketMatch method of Yeturu and Chandra,[6] the SOIPPA method of Xie and Bourne.[22] Each of these algorithms can compute local similarities, and each are capable of producing alignments between protein binding sites. In each case, comparisons were made either by analyzing the performance of PSIM on sets used by the authors of other methods (the Kinnings, Kahraman, Hoffman, and Yeturu sets), or by analyzing the performance of other methods on a set introduced in this study (PocketMatch and SOIPPA on the p38$\alpha$ set). In the former comparisons, PSIM performed as well as the best reported methods, within the ability of statistics to distinguish among them. In the latter comparison, while limited in scope, PSIM pocket similarities exhibited a unique relationship to bound ligand similarities.

Many more approaches for protein structure comparison have been developed, notably the Dali work of the Holm and Sander group[2,23] and the combinatorial extension method from the Shindyalov and Bourne group[3,24,25] These methods have been developed primarily for the study and characterization of the space of protein structures and the relationship of global structure to function, which is a different focus than the proposed work, which seeks to address local comparisons between protein surfaces. Closer to this concept are methods for three-dimensional protein motif identification (e.g. SeqFEATURE[4] and GASPS).[26] These approaches identify local structural features within proteins (e.g. the catalytic triad of serine proteases) that establish a functional motif for proteins with related function. As with the global alignment methods, these have been developed primarily for characterization and annotation of protein function, but not to distinguish, for example, the differences in ligand specificity between two different serine proteases. A very different family of approaches also geared toward protein function annotation describes pockets based on fingerprint vectors, which do not yield alignments. A recent example called FuzCav was reported by Weill and Rognan.[5] The approach constructs a cavity descriptor to characterize a protein binding site based upon the counts of 4833 different pharmacophoric features. The approach is very fast, requiring only comparison of precomputed vectors, and it was effective in distinguishing different classes of proteins. Our approach is geared specifically toward detailed pocket surface comparison based on joint alignment, and we view it as being complementary.

Because of PSIMs focus on surface shape and charge it can isolate small changes between protein structures that might go unnoticed with other methodologies. It is also the only method that shares its underlying formalism with a mature method for computing small molecule molecular similarity.[10, 11]

## 2.5 Conclusions

We have shown that a local, surface-based, protein pocket similarity metric yields informative results across several levels of protein inter-relatedness. Among closely related kinases (including many alternate conformations of single proteins), we showed that the PSIM metric grouped proteins bound to similar ligands more closely than those bound to more divergent small molecules. Among a more diverse set of kinases, we showed that kinase ligand binding specificity was related to our direct computation of protein pocket similarity, with proteins binding the same ligands having more similar pockets to one another (even despite quite different primary sequences) than proteins not sharing ligand binding preferences. Among an extremely diverse set of proteins, all of which bind ATP, we showed that the means by which ATP is bound varies and that different structural strategies can be identified purely based on local surface similarity. The structural motifs were strong enough that methods making use of even multiple sequences were unable to correctly identify the motifs. Among heterogeneous sets of proteins, where protein classes were represented by diverse exemplars (the Kahraman and Hoffman Sets) or by highly similar exemplars (the Yeturu Set), the PSIM approach performed at least as well as the best reported methods. Unique to PSIM was the correspondence between protein pocket similarity and ligand similarity.

Within each of these levels of protein comparison comes the opportunity for future application of the methodology. At the level of highly related proteins (e.g. conformational variants of a particular kinase), automated alignment and selection of conformation variants for molecular docking studies is of interest. Current approaches generally rely on single protein conformations for screening libraries of ligands, but addressing protein conformational variability has clear benefits.[27] At the level of related families of proteins that are interesting from a drug discovery point of view (e.g. serine proteases or human kinases), careful comparison of active sites may help identify potential sources for off-target effects of small molecule therapeutics. Conceivably, nonobvious off-target effects could also be identified, given that sequence relatedness is not required for the method to identify strong structural motifs.

Among the broadest set of proteins, one of the most interesting possibilities lies in functional annotation. Here, there are two clear opportunities. The first combines the structural alignment approach with local sequence motif identification in the hope that the former amplifies the signal for the latter, enabling identification of as yet unknown sequence motifs that could be used to annotate functions for proteins whose structure has yet to be elucidated. The second would seek to comprehensively profile proteins whose structure has been determined specifically to help yield convincing functional information.[28]

There are also significant technical areas in which further study will be required. Understanding the thresholds at which different levels of similarity support some level of confidence in making a functional annotation will require broader study of larger sets of proteins. It is likely that raw similarity values will be context-dependent in the sense that a particular similarity value computed against a large, complex, binding site would probably mean more than the same value computed against a smaller and less complex site. The method is also computationally expensive in its current implementation, requiring on the order of 1 min per comparison on standard desktop hardware. The speed issue derives from the adaptation of this method from small molecules, where alignment optimization involves moving one ligand onto another. Applied in the most straightforward manner, this is inefficient for proteins owing to their large number of atoms. An adaptation of the approach where molecular observer points are moved, with proteins remaining fixed until a final alignment is produced, will yield a substantial speed benefit. Gains in computational throughput will support broader characterization of the PSIM approach and will offer more practical application of the method for prospective studies.

# Chapter 3

# Surflex-Dock: Docking Benchmarks and Real-World Application

Spitzer, Russell, and Ajay N. Jain.

“Surflex-Dock: Docking benchmarks and real-world application.”

*Journal of computer-aided molecular design*

26.6 (2012): 687-699.

## 3.1 Abstract

Benchmarks for molecular docking have historically focused on re-docking the cognate ligand of a well-determined protein-ligand complex to measure geometric pose prediction accuracy, and measurement of virtual screening performance has been focused on increasingly large and diverse sets of target protein structures, cognate ligands, and various types of decoy sets. Here, pose prediction is reported on the Astex Diverse set of 85 protein ligand complexes, and virtual screening performance is reported on the DUD set of 40 protein targets. In both cases, prepared structures of targets and ligands were provided by symposium organizers. The re-prepared data sets yielded results not significantly different than previous reports of Surflex-Dock on the two benchmarks. Minor changes to protein coordinates resulting from complex pre-optimization had large effects on observed performance, highlighting the limitations of cognate ligand re-docking for pose prediction assessment. Docking protocols developed for cross-docking, which address protein flexibility and produce discrete families of predicted poses, produced substantially better performance for pose prediction. Performance on virtual screening performance was shown to benefit by employing and combining multiple screening methods: docking, 2D molecular similarity, and 3D molecular similarity. In addition, use of multiple protein conformations significantly improved screening enrichment.

## 3.2 Introduction

The field of small molecule docking was initiated by the pioneering work of Kuntz and Blaney on rigid ligands in the 1980's.[29] The first practical, flexible, and fully automatic methods began to appear in the 1990's, with AutoDock,[30,31] GOLD,[32,33] Hammerhead,[34–36] and FlexX.[37,38] The earliest efforts typically demonstrated successful re-docking of ligands into their cognate protein binding sites, usually with just a handful of examples, frequently including cases such as trypsin/benzamidine (3PTB), streptavidin/biotin (1STP), and DHFR/methotrexate (4DFR). With the publication of the 1997 GOLD validation paper,[33] reporting pose prediction performance on 100 complexes, the scale of validation experiments for ligand pose prediction changed permanently. Publication of the independent benchmarking of docking algorithms by Rognan's group in 2000 added virtual screening assessment (on thymidine kinase and estrogen receptor) to the types of formal assessments commonly made of docking algorithms.[39] Development of the Surflex-Dock approach (first described in 2003[40]), the descendent of the Hammerhead system, benefited from cognate-docking benchmarks for pose prediction assessment (81 complexes derived from validation of GOLD[33]) and from benchmarks for virtual screening assessment (2 target systems, known positive ligands, and a decoy set from Rognan's group[39]).

The early years of the new millennium saw the introduction and popularization of additional docking algorithms, with independent benchmarking becoming increasingly prevalent. Studies from Perola et al.[41] and Warren et al.[42] were particularly influential. During this same period, larger and more diverse virtual screening benchmarks were developed, notably the set of 29 screening target systems for testing scoring function optimization by Pham and Jain[43] and 40 screening targets forming the DUD set by Huang, Shoichet, and Irwin.[44] With respect to measuring pose prediction, the importance of high-quality structures was gaining prominence, highlighted by the publication in 2007 of the Astex Diverse set of 85 protein ligand complexes.[45] At the same time, the limitations of using cognate ligand re-docking were beginning to be recognized, for example by Sutherland et al.[46] and also by Verdonk et al.[47] who each developed benchmarks for assessment of non-cognate pose prediction.

A special symposium on evaluation of molecular modeling methods took place at the Fall 2007 National ACS meeting, with special attention paid to the issues governing proper assessment of docking algorithms. The meeting yielded several papers, published in a special issue of this Journal, introduced with an editorial by the symposium co-organizers Nicholls and Jain.[48] While consensus among the broader community has been elusive, several issues of central importance were identified relating to benchmark construction and statistical methodology. In the area of virtual screening evaluation, some agreement was made as to sensible statistical methods for measuring enrichment, but decoy set design approaches remained controversial. These consisted of two types: "designed" decoy sets chosen to mimic properties of a set of known actives for a particular target and "agnostic" decoy sets chosen to mimic properties of a typical small molecule screening library. In the area of pose prediction assessment, serious problems with cognate docking benchmarks were highlighted involving "memory effects" that develop when optimizing a protein's pocket structure in the presence of the ligand to be docked as a test.[49]

This paper is part of a collection devoted to a follow-up to the aforementioned symposium that took place in Spring 2011, co-organized by the authors of the lead editorial in this special issue of the Journal of Computer-Aided Molecular Design.[50] Participants were asked to present comparable data and analyses on pose prediction using the Astex Diverse set of 85 protein ligand complexes for pose prediction and on screening utility using the DUD set of 40 protein targets, along with known positive ligands and designed decoy sets for each target. Both sets involved multiple aspects of manual re-curation, especially as to the protein structures themselves.

Performance of Surflex-Dock on the re-prepared Astex85 set was not statistically significantly different than our previous application to the originally released data set,[27] with success rates for single top-scoring poses within 2.0Å RMSD ranging from 66-80% depending on input coordinate variations and run conditions and success rates for best of 20 top-scoring poses of approximately 95%. Performance of Surflex-Dock on the re-prepared DUD40 set yielded a mean ROC area of 0.72 (stdev. 0.15) and mean 1% ROC enrichment of 19 (stdev. 14.5). This was not statistically significantly different than what was reported in the independently published report of Cross et al.,[51] which compared results for several docking methods. They concluded that GLIDE and Surflex-Dock were capable of superior performance in both pose prediction and in virtual screening to the other methods tested: DOCK, FlexX, ICM, and PhDock. Use of SP mode for GLIDE and enabling ring flexibility for Surflex-Dock produced the best overall results in that study.

In addition to the baseline benchmarking that provided a comparative platform for the symposium, we addressed four additional questions, two related to pose prediction and two related to virtual screening: 1) to what extent are subtle changes in protein preparation capable of yielding large improvements in nominal pose prediction performance? 2) is it possible to make use of protein pocket adaptation *during* the docking process to produce high quality pose prediction results? 3) is a multipronged strategy for virtual screening, which combines docking, 2D similarity, and 3D similarity, more robust and reliable that one method alone? 4) is it possible to make use of multiple protein conformational alternatives to improve virtual screening performance without requiring *ad hoc* scoring adjustments?

Use of multiple alternative protein conformations was also shown to have a significant positive impact in two target systems where data were available to make direct comparisons. For certain classes of proteins, flexibility in the protein binding site can test the limits of the assumption of protein conformational rigidity in single-structure docking. Kinases, for example, present varying conformations which alter the binding site volume and geometry. Docking to multiple conformations that represent the variation present in a protein binding site, allows a docker to recover from the limitations of a single structure strategy. Use of multiple protein conformations is made possible through the use of Surflex-PSIM (Protein SIMilarity), a technology for the alignment and comparison of protein binding sites. Given a set of crystal structures for a docking target, PSIM provides a fast and accurate method of aligning and selecting structures which best represent the diversity of the pocket. The multi-structure docking protocol was able to provide large gains over a traditional single-structure docking protocol in two systems.

We observed gains in pose prediction success rates of nearly 20 percentage points by making very small changes to protein structures (typically 0.3Å RMSD within the protein pocket) *prior* to docking by joint optimization of protein and cognate ligand. However, we also showed that very high success rates could be obtained using a practical procedure that adapted protein pockets *during* the docking process and produced *pose families* based on clustering and a Boltzmann weighting scheme. With respect to virtual screening, we showed that using the combination of docking and similarity approaches produced robust performance, with early enrichment of 15-fold or greater 75% of the time and overall ROC area of 0.80 or greater 60% of the time. Use of multiple alternative protein conformations was also shown to have a significant positive impact in two target systems where data were available to make direct comparisons.

## 3.3 Data and Methods

The primary molecular data sets for this study were obtained as part of participation in a symposium. The details of the pose prediction set, 85 complexes adapted from the Astex Diverse Set,[45] which will be referred to as the Astex85 set, can be found in the lead editorial of this special issue.[50] The details of the virtual screening benchmark set, 40 targets along with nominal true ligands and designed matched decoy sets was adapted from the DUD benchmark,[44] which will be referred to as the DUD40 set, can also be found in the lead editorial. For both benchmarks, substantial re-preparation of protein structures was carried out in order to provide a common set of coordinates (including hydrogen atoms) to participants. Modifications to ligand structures were quite significant for the Astex85 set, where fresh input coordinates were generated in order to fully eliminate memory effects of bound cognate ligand poses. For the DUD40 set, some targets received a degree of re-curation of positive examples of ligand structures (e.g. to address bond order errors in trypsin ligands and chirality errors in PDE5 ligands).

All docking and similarity calculations were carried out using standard protocols with Surflex-Dock and Surflex-Sim version 2.514. For pose prediction tasks on the Astex85 set, the "-pgeom" parameter set selecting the geometric docking search protocol was used, with "+ring" additionally since the input ligands coordinates often had strained ring systems. The limited protein pocket adaptation tasks were carried out using standard docking followed by pocket optimization (the "rescore_multi" command) and pose family generation (the "posefam" command) as reported in the paper introducing the Surflex-Dock pocket optimization protocol.[27] Demonstration of the effects of protein structure pre-optimization for the Astex85 set was carried out as previously illustrated on a different set initially reported by Vertex.[41,49,52] For virtual screening tasks on the DUD40 set, the "-pscreen" parameter set selecting the fast screening search protocol was used, and ring search was not enabled. Comparisons were also made using Surflex-Sim's 3D surfaced-based molecular similarity approach and the Surflex 2D approach called GSIM.[11,53,54] For application of molecular similarity, the cognate ligand of each protein target in question was used as the target of the similarity search. We also carried out tests of the Surflex-Dock multi-structure docking protocol (the "mdock_list" command[27]) using standard screening parameters.

Data archives can be requested through jainlab.org.

Table 3.1: Summary of results for pose prediction accuracy on Astex Diverse Set of 85 complexes.

| | % Correct: Top | % Correct: Best | Mean Top RMSD (stdev) |
|---|---|---|---|
| **Original Astex85** | 80 | 96 | 1.66 (1.82) |
| **Re-prepared** | 66 | 93 | 2.18 (2.09) |
| **Proton-optimized** | 73 | 95 | 1.85 (1.88) |
| **All-optimized** | 84 | 95 | 1.34 (1.46) |
| **Top pose family** | 68 | - | 1.99 (2.19) |
| **Top two pose families** | 82 | - | 1.31 (1.56) |
| **All pose families** | 87 | - | 1.15 (1.37) |

## 3.4 Results and Discussion

The performance of Surflex-Dock on the Astex85 and DUD40 sets has been published previously. For the former set, using the original structures from Hartshorn et al.,[45] we used the set to draw a contrast between the ease of cognate ligand re-docking compared with non-cognate docking.[27] For the latter set, a careful and comprehensive comparison of several docking programs was carried out by Cross et al.[51] Both studies were relatively recent and made use of up-to-date Surflex-Dock versions. The modifications to the benchmarks for this symposium were not extensive, so those published results differed little from what is reported here. In what follows, first for pose prediction and then for virtual screening efficiency, we will briefly summarize the baseline results while highlighting differences from previous work. In addition, we will address questions involving protein-ligand complex pre-optimization, protein pocket adaptation as part of the docking process, use of hybrid screening approaches that combine docking and similarity computations, and use of multiple protein structural examples as the target of virtual screening.

### 3.4.1 Cognate Docking: Performance on the Astex85 Set

Table 3.1 reports the results of several docking runs on the Astex85 set, making use of different protein and ligand starting coordinates or run protocols. The top two rows are directly comparable. The first row shows results on the protein and ligand coordinates released by the originators of the Astex85 set;[45] these results had been reported as part of a study exploring the effects of protein conformational adaptation.[27] The second row shows results on the re-prepared Astex85 set.[50] The key differences in the protein coordinates stemmed from fresh structure refinement in the re-prepared set and optimization of proton positions using GoldScore with the cognate ligand in the original set. The key differences in the ligand coordinates stemmed from use of CORINA to produce fully agnostic memory-free ligand starting coordinates in the re-prepared set and a protocol of torsional and alignment randomization and minimization for the original set. The differences in protein coordinates yielded relatively subtle changes in the energetic landscape to be probed by ligand docking. The differences in ligand coordinates were more profound in many cases, resulting in important changes in protonation state, tautomeric state, and input ring conformations. The docking success rates (proportion of dockings with RMSD $\leq 2.0$Å) were somewhat better for the original set than for the re-prepared set (80% vs. 66% for top scoring pose and 96% vs. 93%). However, neither the success rates, nor the mean RMSD values, differed in a statistically significant manner. Figure 3.1 shows the comparison of the cumulative histograms.

There are two key reasons that cognate ligand re-docking is an artificial test of pose prediction. First, this is never the operationally important application in a real-world use-case of docking for binding-mode prediction. In a real-world application, a modeler would choose to explore the binding mode possibilities for some ligand that is different from any whose bound configuration is known. Depending on the protein,

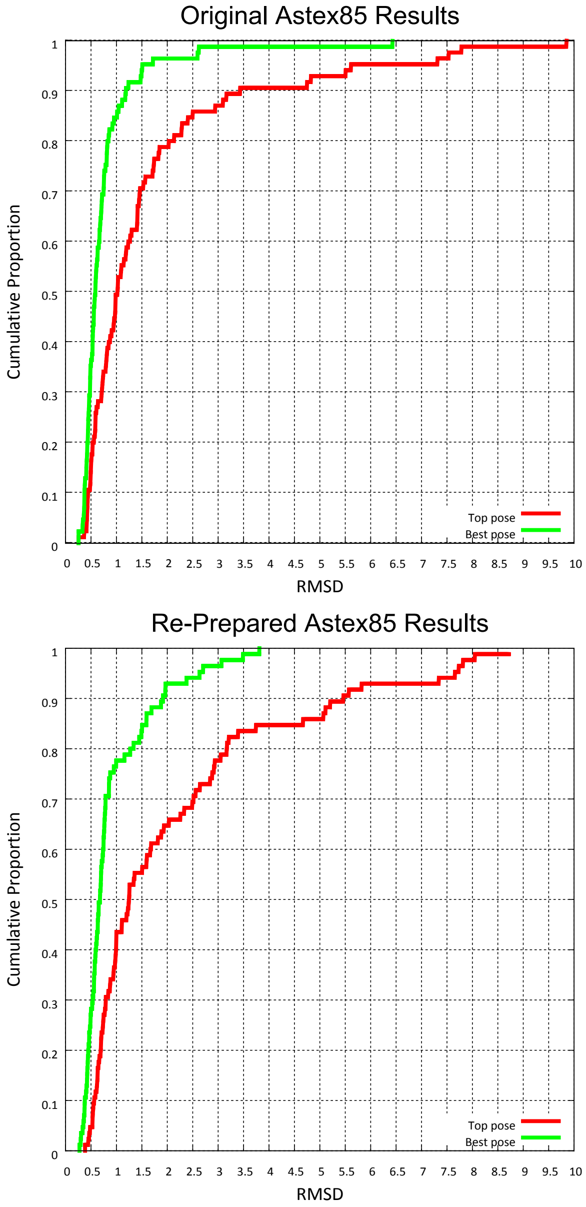


Figure 3.1: Comparison of results from the original Astex Diverse Set release compared with the Re-prepared Set. Differences in top-scoring pose performance were larger than for best pose of top 20, but were not statistically significant at the 2.0Å success cutoff.

the binding pocket may adapt in subtle or unsubtle ways, but it will generally change *at least a little*. So, the "memory" of the ligand expressed in the particular pocket coordinates of a protein used in a cognate docking test represent an advantage that is never present in a real-world application. Second, as we have shown before,[49, 52] very small changes in protein pocket conformation, even involving just proton movement, can have a large impact on pose ranking *within* the top set of docking poses returned. In particular, coordinate optimization of a complex can exacerbate the memory effect already present in the cognate protein structure. To illustrate the magnitude of this effect, beginning with the re-prepared Astex85 set, for each complex, we performed joint optimization of ligand and protein binding pocket, either only for protons or including non-hydrogen protein pocket atoms as well.

We then repeated the docking computations using these protein variants. These results are summarized in the middle rows of Table 3.1. While little effect was seen on success rates for best pose among the top 20, a nearly 20-point increase in success percentage for the top pose was obtained using the protein variant generated with all-atom pocket optimization. The difference between 66% and 84% success rates for 85 complexes was statistically significant (p = 0.01 using Fisher's exact test). Figure 3.2 shows the corresponding cumulative histograms of observed RMS deviations. The red curves correspond to the unmodified re-prepared protein coordinates (as in Figure 3.1). The *only* difference between the red and green curves was changes in proton positions for the latter, and the blue curve shows the effects of allowing non-hydrogen atoms as well to adapt to a local minimum prior to docking. The magenta curves make one additional change: measuring the RMS deviation from the *optimized* cognate ligand coordinates (for the all atom protocol) instead of the crystallographically modeled ones. Complex pre-optimization has a very significant impact on top-scoring pose performance, owing to the enhancement of the particular local minimum corresponding to the known bound ligand configuration. This effect

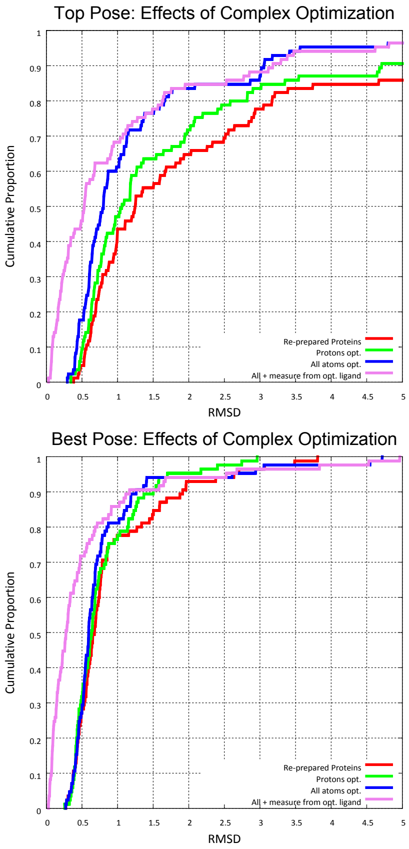


Figure 3.2: Optimization of protein-ligand complex prior to docking can have a significant impact on nominal pose prediction performance, especially for top-scoring pose. The two graphs show Surflex-Dock performance on the re-prepared Astex85 set under different preparation and RMSD measurement protocols (the top graph shows top pose RMSD cumulative histograms and the bottom shows corresponding information for best pose of the top 20 returned). Results on the re-prepared Astex85 set are shown in red, results for the same proteins after proton optimization in green, and all pocket atom optimization in blue. The magenta lines shows the change in RMSD when measuring deviation from the optimized ligand coordinates rather than the crystallographic coordinates for all atom pocket pre-optimization.

derives from very small movements in protein atoms (see Figure 3.3). The effect of measuring from the optimized ligand coordinates has an enormous impact on the fraction of very low RMSD results, which also skews statistics involving mean RMSD. Given the uncertainty in coordinate precision for even high-quality structural models, high proportions of RMS deviation values for pose prediction less than 0.5Å suggest this type of coordinate optimization.

This effect has been discussed more extensively in trying to understand differences in nominal pose prediction performance among docking methods with different congruence to an energy function used for protein optimization.[41,52] It has also been discussed in the context of the appropriateness of protein optimization and RMS deviation measurement from optimized ligand coordinates, as practiced by some methods

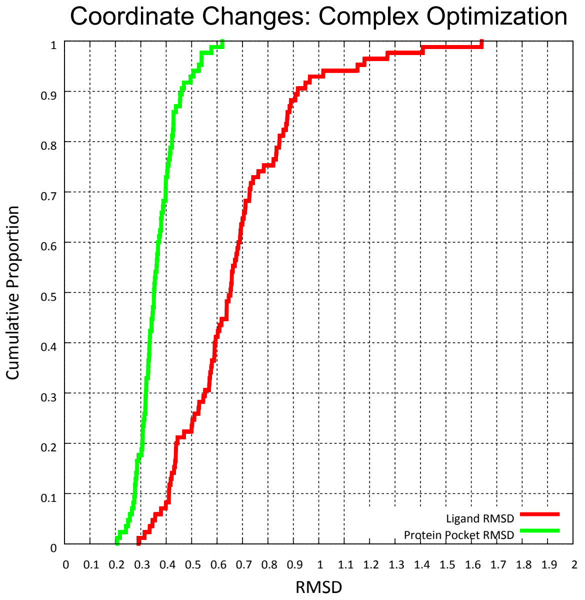
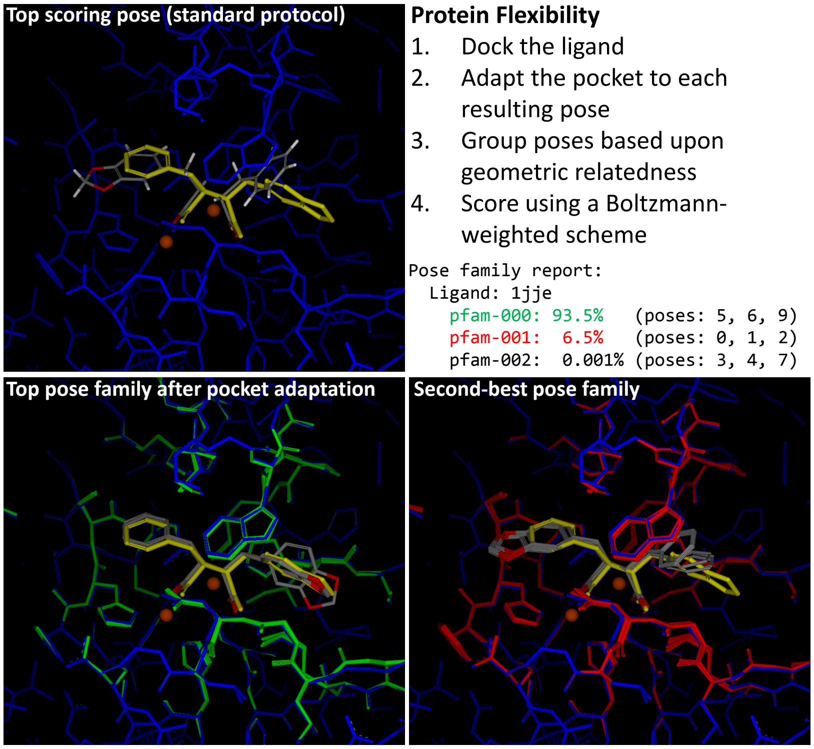


Figure 3.3: Coordinate changes were small, especially for the protein, even with all-atom coordinate optimization of complexes. These tiny coordinate changes gave rise to significant changes in the pose ranking that result from docking to the modified proteins.

developers,[55] in a paper devoted to questions involving docking method evaluation.[49] We do not believe that such protocols produce performance estimates that will reflect real-world application of docking methods.

Unbiased Protein Atomic Movement: As we have seen, pre-optimization of protein coordinates using an energy function that is congruent to the one being used in a docking system can predispose performance results very favorably. We believe that the best approach to avoid such problems is to test pose prediction on non-cognate ligands, often termed cross-docking. We have previously shown substantial improvements on a challenging cross-docking benchmark using Surflex-Dock's multi-protein docking protocol coupled with protein pocket adaptation and a pose clustering and rescoring technique that yields pose *families*.[27] To illustrate the effects of this protocol in the context of data available for the symposium from which this paper resulted, we applied it to the Astex85 set. In the full protocol, protein conformational variants representing large movements are used, but to illustrate the effects of pocket adaptation on the Astex85 set, only the single re-prepared protein structure for each complex was used.

Figure 3.4 illustrates the procedure with the test complex corresponding to PDB code 1JJE. In this example, the top scoring pose from Surflex-Dock using the standard protocol was incorrect, shown in atom color at top-left along with the correct pose shown in yellow. The ligand is partially symmetric, and the top-ranked pose is a flip that places the metal-interacting moieties correctly. The Surflex-Dock pocket adaptation protocol optimizes the final docking poses within the protein pocket while allowing the pocket atoms to move, subject to a covalent force-field as well as inter-molecular scoring energy terms that govern the docking. To enhance sampling, multiple small perturbations may be carried out for each pose (in this case, two perturbation were used). A score that represents the overall energy of ligand, protein pocket, and their non-bonded interactions (with each other and among themselves) is computed for each jointly optimized configuration. The resulting ligand poses are clustered based on RMSD, and a Boltzmann-based formula is used to apportion percentages to each such pose family, with families that have too low a percentage eliminated from the output. For the example in Figure 3.3, there were three families


Top scoring pose (standard protocol)
Top pose family after pocket adaptation
Second-best pose family

**Protein Flexibility**

1. Dock the ligand
2. Adapt the pocket to each resulting pose
3. Group poses based upon geometric relatedness
4. Score using a Boltzmann-weighted scheme

```
Pose family report:
  Ligand: 1jje
    pfam-000: 93.5%   (poses: 5, 6, 9)
    pfam-001:  6.5%   (poses: 0, 1, 2)
    pfam-002:  0.001% (poses: 3, 4, 7)
```

Figure 3.4: Use of protein pocket optimization and pose family generation offers a means to explore changes in protein pocket configuration on ligand binding in a way that is not biased. The top-left panel shows the single top scoring pose (atom color) for test case 1JJe, which was a flip of the crystallographic pose (yellow). The top scoring pose family (bottom left, atom color) was correct, resulting from rescoring after jointly optimizing the docked ligand poses, which resulted in some protein movement (green). The second ranked pose family (bottom right) required slightly more alteration of the protein binding pocket, especially at the left-hand side (red).

generated, with the top family accounting for 93.5% of the expected joint configurations, the second family 6.5%, and the last one just 0.001%. By taking into account the overall energetics of the complex, the top family (bottom left) now clearly contains the experimentally determined pose. The original ranks of the poses that gave rise to the top family were 5, 6, and 9. The second ranked family arose from the top 3 original poses, and shows the flipped orientation of the ligand. The atomic movements of the protein (green for the top family and red for the second-ranked one) were small, but sufficient to produce the correct ranking.

An advantage of this procedure is that one gains some degree of information as to the uncertainty in the pose prediction. This is reflected in the amount of movement exhibited by the ligand within each pose family and also by the number of pose families produced. Figure 3.6 shows an example (PDB code 1SJ0) where there was a flexible ring system in the ligand in question. The ligand coordinates used as input for docking contained a reasonable ring conformation, but it was incompatible with the correct binding mode. The middle panel shows the resulting docking *without* ring search for illustration only. With ring search enabled (as it was for the primary results for pose prediction), and making use of the pocket adaptation procedure, a *single* pose family was produced (bottom panel of Figure 3.5), which clearly encompassed the correct binding mode. The pose within the family that had the smallest RMS deviation was within 0.6Å of the experimentally determined ligand coordinates. A single pose family was generated for 20 of the 85 complexes. For this group, the mean RMSD was $0.77 \pm 0.62$, with 95% (19/20) having RMSD $\leq 2.0$Å. The bottom rows of Table 1 summarize results for the pocket adaptation protocol. The top ranked pose family produced just a marginal improvement over the original docking protocol, but by considering the top two families, the success rate improved from 66% to 82%. When considering all pose families that were produced, we observed a success rate of 87% ($p \ll 0.01$, compared with 66% success by Fisher's exact test). Figure 3.6 shows the cumulative histograms for the top and top-two pose family results. Without relying on the fortune of well-oriented protein pocket hydrogen atoms, we observed

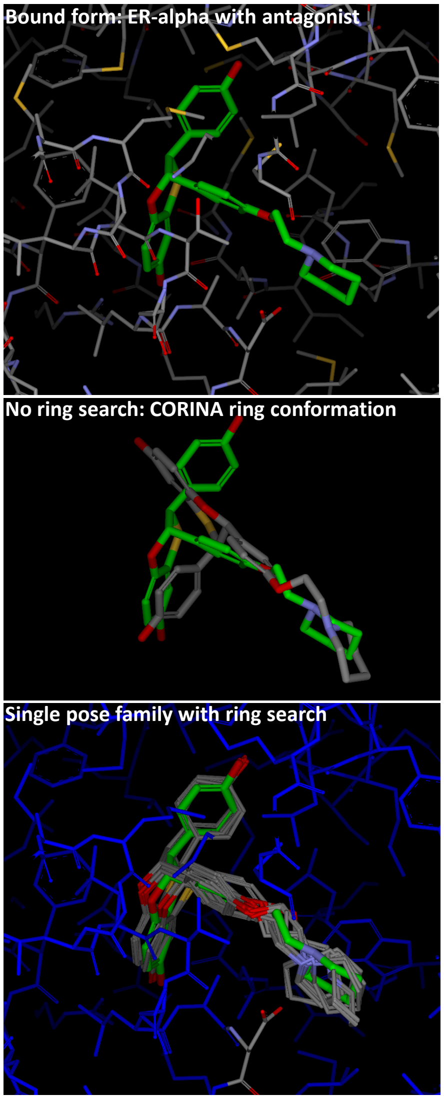

Bound form: ER-alpha with antagonist
No ring search: CORINA ring conformation
Single pose family with ring search

Figure 3.5: The combination of Surflex-Dock's ring search and pocket adaptation and pose family protocol rescued a poor result using the given ligand coordinates (middle panel), yielding a single pose family (bottom panel), which closely covered the crystallographically determined ligand pose (green). Cases where only a single pose family were generated yielded correct results 95% of the time (see text for details).

very strong results, especially for the two pose family case, but even in the single family case, there were significant improvements at low RMSD.

Clearly, results for a single top-ranked pose and those produced when considering multiple families are of a different type. However, we believe that the modeling question addressed by pose prediction with docking is better matched to examining

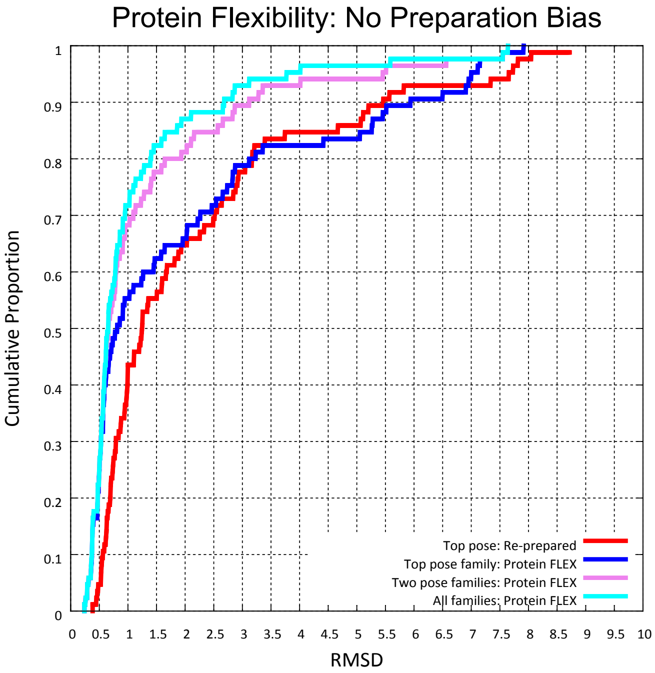


Figure 3.6: Results applying protein flexibility *during* the docking process show an improvement when considering only the top scoring pose family (blue) over results on the re-prepared set using the standard docking protocol (red), especially at low RMSD values. The gain obtained by looking at the top two pose families was very substantial (magenta line), reflecting the common occurrence of "flips" of pseudo-symmetric ligands that received very close scores. Consideration of all pose families generated (cyan line) yielded a further small gain. More than 90% cases produced five or fewer families.

a small number of pose families, each associated with a percentage of coverage, than it is to examining a single pose. For the Astex85 set, 24% of the cases produced a single family, 45% produced two or fewer families, 68% three or fewer, and less than 10% produced more than five (with a maximum of seven). The type of alternative flip shown in Figure 3.4, where the ligand is pseudo-symmetric and where both orientations appear plausible, represent the most common variations among the different pose families. Typically, a key interaction is common among the different alterna-

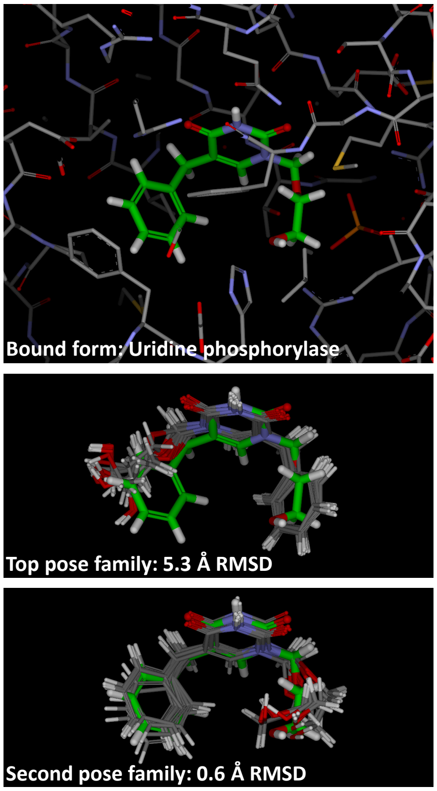

Bound form: Uridine phosphorylase
Top pose family: 5.3 Å RMSD
Second pose family: 0.6 Å RMSD

Figure 3.7: PDB code 1U1C was a test case where ligand density was poor (the top panel shows the nominal bound configuration). The middle panel shows the top scoring pose family, which was a flipped orientation around the central symmetric ring system relative to the second ranked pose family (bottom panel). It was not clear whether the crystallographic data could reliably distinguish these two alternatives.

tives, with the pose families stimulating development of new hypotheses for where interaction opportunities may exist. There were also examples, as with PDB code 1U1C, where the complexes were highlighted as having poor ligand density. This case is depicted in Figure 3.7. The top-scoring pose family was “incorrect” (middle panel) but the second-ranked one was “correct” (bottom panel). The two alternatives were symmetrically flipped around a central axis, but the alternatives may not be definitively distinguished by the crystallographic data. Even if the nominally correct pose represents a truly better energetic configuration in the real biological system, we believe that the alternate binding mode is informative, suggesting the potential for hydrophilic interactions with the left-hand side of the pocket.

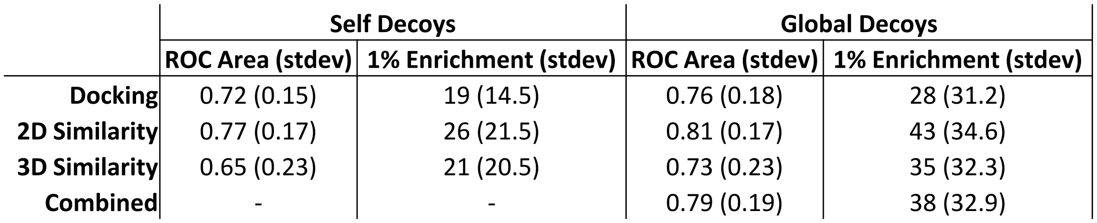
Summary: Our analysis of performance of Surflex-Dock on the Astex85 set makes four primary points. First, the minor differences in set preparation made little difference in pose prediction performance, with the bulk of the difference probably arising from the use of a different randomization protocol for ligand starting configuration in the re-prepared set compared with the original set. Second, cognate re-docking of ligands as a means to test pose prediction is fraught with difficulty, since it is so dependent on the congruence between the means used for protein structure preparation and the method to be used for docking. Use of coordinate optimization schemes that make small and benign changes to protein coordinates can produce very significant changes in the ranking of poses whose energies are close. Third, since it is clearly necessary to address protein atomic movement in order to produce useable results for cross-docking, continued use of rigid protein cognate-docking tests is difficult to support. Last, judging performance based on the deviation between the single top-ranked pose and the experimentally determined one is much less informative than considering some form of pose clustering. Such techniques usually yield few distinct solutions, the vast majority of which are reasonable, with the number of solutions related to the confidence in pose prediction.

### 3.4.2 Virtual Screening: Performance on the DUD40 Set

It is useful to place the development of the DUD set in context. Introduced in 2007,[44] it was meant to address two significant problems in assessment of docking for the purpose of virtual screening. First, issues had been raised with respect to the physical characteristics of decoy sets and the ease with which one could distinguish active ligands from such decoys. Notably, the set from Rognan's group[39] was characterized by many, not unfairly, as being too hydrophobic compared with drug-like compounds. Second, other virtual screening benchmarks had either limited numbers of active ligands for each target or had limited numbers of targets, or both. The largest such set at the time was that from Pham and Jain,[43] consisting of 29 targets, but with a maximum of 20 ligands per target. That report included two decoy sets: the Rognan set (990 molecules) as well as one derived from screening molecules meant to have similar properties to drug leads (1000 molecules). The DUD set had more targets (40), more active ligands per target (an average of about 70), and a design-based approach to constructing decoys. For each target, the idea was to come up with 40 decoys per active ligand that replicated aspects of physical characteristics but avoided 2D molecular similarity to any of the known actives. Experiments using DOCK were carried out with all ligands and decoys against all 40 protein targets. Comparisons were also made between the DUD decoys and other decoy sets, with the largest differences existing between the amalgamated (or global) DUD set of 95,316 decoys and the Rognan set. It is worth noting that the authors of the DUD set advocated using *both* the "own decoys" (here termed "self decoys") and "amalgamated DUD" (here termed "global decoys'), since they represent different challenges.[44]

Table 3.2 summarizes results for virtual screening assessment on the DUD40 set, both using the self decoys and the global decoys for Surflex-Dock, Surflex-Sim 3D molecular similarity, and for the GSIM 2D similarity metric. The results we obtained for Surflex-Dock using the DUD self decoy set on the re-prepared DUD40 set did not differ from those reported recently by Cross et al. on the original DUD set.[51] One striking feature was that while ROC AUC did not change much when comparing self decoy results to global decoy results (typical shifts in mean AUC of just 0.04), the typical early enrichments (measured using ROC 1% true-positive to false-positive ratios) nearly *doubled.* Data are presented in more detail in Figure 3.8, showing the dramatic improvement in early enrichment when using the global decoy set (top two graphs). For both docking and 3D similarity, early enrichments of greater than 45-fold were observed in roughly one-third of cases. It is also important to note that simple 2D similarity searching performed very well owing to the common occurrence extreme topological similarity of DUD actives to be retrieved compared with the cognate ligand of the protein structural target. This issue has been analyzed in greater detail previously,[54] especially concerning the use of DUD in evaluating molecular similarity methods. To better approximate the real-world application of virtual screening, we also evaluated the performance of the combination of docking, 2D, and 3D similarity. Information from the three methods was combined by computing the product of the resulting ranks for each ligand. On a target-by-target basis, the hybrid approach was always better than the worst of the individual approaches, with mean improvement in AUC of 0.13 ± 0.08. The hybrid approach was generally slightly worse than the *best* of the individual approaches, with mean decrease in AUC 0.07 ± 0.11. Notably, the

Table 3.2: Summary of results for virtual screening performance on DUD set of 40 targets.



| | Self Decoys | | Global Decoys | |
|---|---|---|---|---|
| | **ROC Area (stdev)** | **1% Enrichment (stdev)** | **ROC Area (stdev)** | **1% Enrichment (stdev)** |
| **Docking** | 0.72 (0.15) | 19 (14.5) | 0.76 (0.18) | 28 (31.2) |
| **2D Similarity** | 0.77 (0.17) | 26 (21.5) | 0.81 (0.17) | 43 (34.6) |
| **3D Similarity** | 0.65 (0.23) | 21 (20.5) | 0.73 (0.23) | 35 (32.3) |
| **Combined** | - | - | 0.79 (0.19) | 38 (32.9) |

hybrid approach *never* performed worse than the most poorly performing individual technique, but it performed slightly better than the best individual technique nearly 20% of the time.

One other aspect of note in Figure 3.8 is that the performance of Surflex-Dock on the Pham/Jain screening set was significantly better than on the DUD40 set. ROC AUC was greater than 0.80 about 75% of the time for the former set compared with just 40% of the time for the latter. This was also reflected in early enrichment rates, with early enrichment of 20-fold or better in 80% of cases for the Pham/Jain set[43] compared with less than 40% of cases for the DUD40 set. In order to understand these differences, we compared the active and decoy structures for each target to ligands bound to those same targets whose structures were available in the PDB. Figure 3.9 highlights the risks involved in designing decoys to look similar to known actives. In the top case, one of the GART decoys is shown in an experimentally determined co-crystal structure with GART. Many of the GART decoys were trivial analogs of the ligand in the 1CDE structure, and it is likely that many of those molecules have reasonable affinity for the GART protein. Similarly, the thymidine kinase decoys include one where an epimer is known to bind TK. While it may be the case that the epimer that was present in the decoy data set does not bind TK at all, we believe this to be unlikely. Further, the extreme similarity of many of the nominal TK decoys to known active TK ligands is of concern. We believe that a very significant portion of the difference between early enrichment performance when comparing DUD self to DUD global decoys stems from "false false positives" as shown in Figure 3.9. At best, such decoys blur the line between potency prediction, where distinctions of 1 kcal/mol are important, and virtual screening, where the expectation is to distinguish larger energy differences. In our comparison of DUD actives to the known, bound configurations that could be found in the PDB, we observed a nearly 1 in 6 rate of unrecoverable structural variations, where changes in chirality were present or bond order variations existed that were not due to tautomerism. We believe that such errors help explain the difference in overall ROC area between the DUD40 set and the Pham/Jain set.

Multiple Protein Structures: In keeping with the idea of trying to ascertain real-world performance, we made a limited attempt to test the effects of using multiple alternative protein structures as the target of virtual screening. To the degree that multiple structures are available for a target that exhibits active site mobility, many modelers would try to take advantage of the additional data. In the case of PDE5 (where active ligand structures had been corrected by the symposium organizers prior to release), there was a large improvement: from ROC area of 0.72 ± 0.06 with a single structure to 0.83 ± 0.06 with four (95% confidence intervals just barely overlapping). The three additional structures (PDB codes 1T9S, 1TBF, and 1XOZ) were chosen and aligned based on a recently published pocket similarity computation patterned after the Surflex-Sim approach.[56] Figure 3.10 shows the primary driver of the improvement: a large positive shift in the scores of the active ligands. The left panel shows a nearly 2 log unit increase in the scores of the 51 known active ligands. The middle panel shows the docked pose of one such active molecule (ZINC04199926 shown in yellow) compared to the bound pose of tadalafil (green). The single original DUD target structure was of PDE5 bound to vardenafil, but the ligand in question has a binding mode much more compatible with the active site rearrangement of PDE5 when bound to tadalafil. The rightmost panel shows the poor predicted poses (shown in red) of the yellow molecule from the middle panel resulting from docking to only the vardenafil-bound PDE5 structure. The poses were clearly wrong, and the scores were much lower than for those making use of the four alternate protein structures, none of which were bound to a ligand sharing the scaffold of the yellow molecule.

In considering making a broader evaluation of this approach, difficulties with curated structures of known active ligands within the DUD40 set presented a serious obstacle. For example, in the case of progesterone receptor (PR), where we expected to see benefits due to rearrangements of the ligand binding domain on binding agonists compared to antagonists, application of the same approach as just shown for PDE5 yielded no improvement: original ROC area of 0.48 and a multi-structure ROC area of 0.46, both indistinguishable from random performance. Of the 27 active ligands, 8 had steroid cores. Of these, 6 were clearly wrong in terms of the chiral configuration of the steroid core. We re-curated a set of 11 active ligands for PR from the PDB, taking care to regenerate the ligand structures from SMILES to avoid any memory of bound configurations. Using the single original DUD protein target structure (with global decoys), we obtained an ROC area of 0.52 ± 0.19. Using three additional structures (1SQN, 2OVM, and 3G8N) chosen as with PDE5, we obtained 0.87 ± 0.10, a clearly significant improvement. An example of this improvement is shown in Figure 3.11.

Use of WOMBAT curated active ligands, which were made available for several targets, did not yield significant performance changes using the standard protocol. No attempt was made to assess error rates in structures within that set.

Summary: Virtual screening using molecular docking is clearly still a significant computational challenge, with highly variable performance depending on the target in question. We have shown that a combination of docking, 2D, and 3D molecular similarity is an attractive approach, exhibiting performance close to the best of any individual method and reliably better performance than the poorest. This approach can be applied to any methodology that produces a ranked list of ligands. Preliminary results indicate that use of multiple target structures can produce marked improvements in screening effectiveness.

Construction of quality benchmarks with numerous targets of diverse character is a serious challenge. We believe that the risks of "designed" decoy sets far outweigh the potential benefits of agnostic sets built to mimic lead-like screening libraries. In particular, presence of decoys that are, in fact true ligands, or whose distinction from being true ligands is a subtle difference in binding energy, artificially decreases estimates of early enrichment. Curation of active ligands must also be done with care. While it may be reasonable for docking systems to begin to cope with internal generation of tautomers or protonation states for ligands, it is not reasonable to expect frank structural errors to be corrected in any fashion by a docking algorithm. Such errors can depress overall ROC AUC values, and they can mask the true effects of algorithm modifications, such as we demonstrated with multi-structure docking.

## 3.5 Conclusions

The field of docking is mature enough to move beyond cognate ligand re-docking, which was introduced more than twenty years ago, as a means to test pose prediction accuracy. Certainly, sets such as the Astex85 set form important resources for methods developers, especially in establishing the baseline feasibility of a new technique. However, cognate docking does not replicate the real-world scenario that is relevant to pose prediction: the case where a new ligand is sufficiently different from the structure of one whose bound configuration is known that a skilled modeler has a serious question about potential binding modes. There are well-curated public benchmarks that address this problem in various degrees of difficulty,[27,46,47] and docking researchers should make an active effort to move away from cognate ligand re-docking.

Data resources to support construction of well-curated benchmarks for measurement of virtual screening performance have evolved to allow for significant improvement over the currently available set of benchmarks. Resources such as BindingDB and PubChem in particular offer well-curated ligand structure and activity data.[57,58] With the ascendance of sophisticated 3D molecular similarity methods as serious alternatives or adjuncts to docking, both for pose prediction and for virtual screening,[10,11,14,53,54,59,60] it is increasingly important to develop such benchmarks. In particular the diversity ligands should be high, and the binding affinities should be typical of verified hits from physical high-throughput screening campaigns.

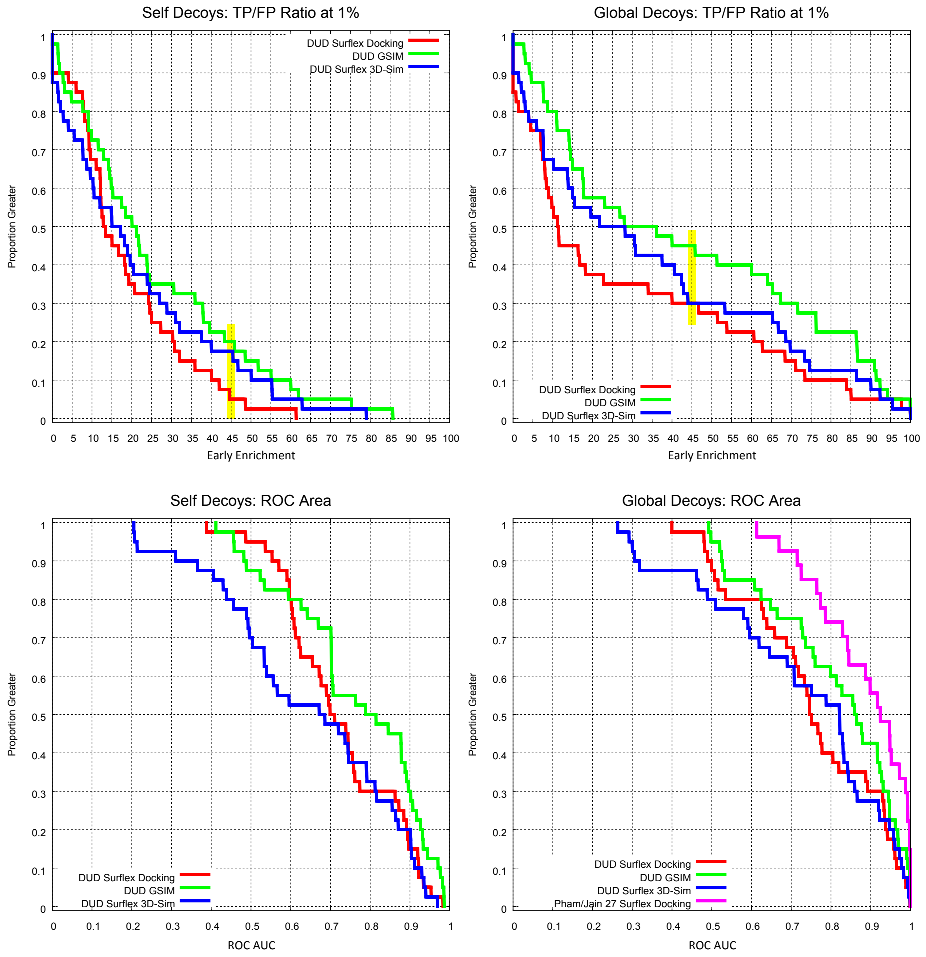


Figure 3.8: Performance on DUD virtual screening benchmark using self decoys and global decoys. The top two graphs show 1% ROC enrichment performance for docking, 3D molecular similarity, and 2D molecular similarity. The bottom two graphs show the corresponding ROC areas, with the bottom right graph also including a comparison to results from a previous study on 27 screening targets. Marked differences in early enrichment (highlighted in yellow at the 45-fold level) were observed, with performance on the global decoys very significantly better. Overall ROC areas changed much less, but the DUD40 set produced poorer docking performance than the Pham/Jain set.

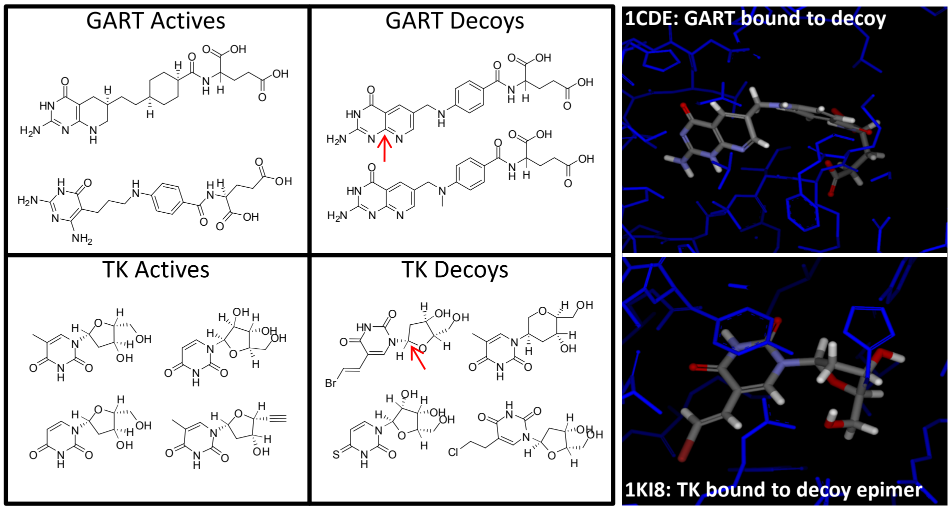


Figure 3.9: Use of designed decoy data sets, which attempt to mimic properties of active ligands bears the risk of inclusion of active molecules within decoy pools. Typical examples of actives and self decoys from the DUD40 set for GART and TK are shown. In the case of GART, the top decoy is shown crystallized with the enzyme that it is not supposed to inhibit (top right). In the case of TK, an epimer of the top left decoy is shown crystallized with TK.

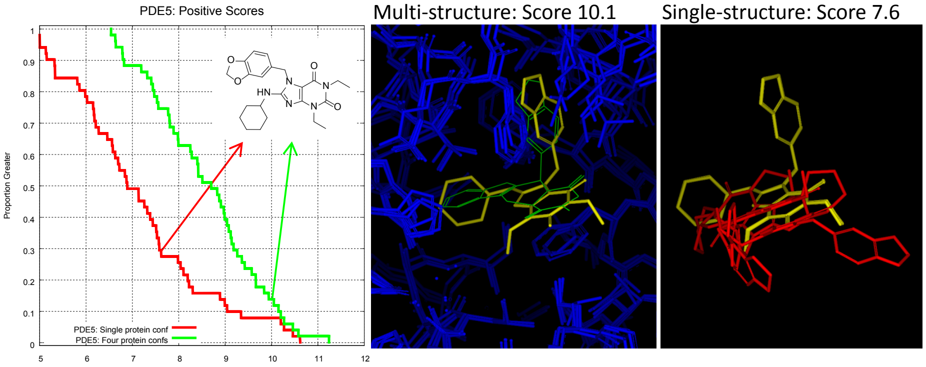


Figure 3.10: Using multiple target protein structures can be helpful, especially in cases like PDE5, where significant rearrangement can occur. The positive ligand ZINC4199926 (shown in 2D at left) scored 10.1 in the multi-structure docking but just 7.6 in the single-structure protocol. Its predicted pose from the multi-structure protocol (middle, shown in yellow) shows a clear relationship to the related PDE5 inhibitor tadalafil (green). Poses from the single-structure protocol (right, shown in red) were very different and likely incorrect. In this case, ROC area improved from 0.72 ± 0.06 to 0.83 ± 0.06 (95% confidence intervals) using the multi-structure protocol instead of single-structure.

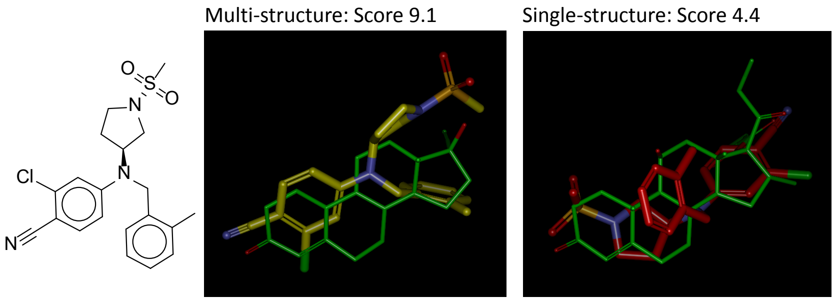


Figure 3.11: A non-steroidal progesterone partial agonist (2D shown at left) was docked with a high score in the multi-structure protocol (middle, shown in yellow carbons). The predicted pose was close to correct (PDB code 3KBA, not shown). It was docked with a low-scoring and incorrect pose in the single-structure protocol (right, shown in red carbons). In both panels, a crystallographically determined steroid structure is shown in green to provide binding-site context.

# Chapter 4

# Protein Function Annotation By Local Binding Site Surface Similarity

Spitzer, Russell, Rocco Varela, Ann E. Cleves, and Ajay N. Jain.

"Protein Function Annotation By Local Binding Site Surface Similarity"

## 4.1 Abstract

Hundreds of protein crystal structures exist for proteins whose function cannot be confidently determined from sequence similarity. Surflex-PSIM, a previously reported surface-based protein similarity algorithm, provides an alternative method for hypothesizing function for such proteins. The method now supports fully automatic binding site detection and is fast enough to screen comprehensive databases of protein binding sites. The binding site detection methodology was validated on apo/holo cognate protein pairs, correctly identifying 91% of ligand binding sites in holo structures and 88% in apo structures where corresponding sites existed. For correctly detected apo binding sites, the cognate holo site was the most similar binding site 87% of the time. PSIM was used to screen a set of proteins that had poorly characterized functions at the time of crystallization, but were later biochemically annotated. Using a fully automated protocol, this set of 8 proteins was screened against approximately 60,000 ligand binding sites from the PDB. PSIM correctly identified functional matches that pre-dated query protein biochemical annotation for five out of the eight query proteins. A panel of twelve currently unannotated proteins was also screened, resulting in a large number of statistically significant binding site matches, some of which suggest likely functions for the poorly characterized proteins.

## 4.2 Introduction

We have previously shown that a local, surface-based, protein binding site similarity metric can identify biochemically relevant relationships between proteins having varying levels of sequence similarity.[56] PSIM distinguished subtle differences between proteins sharing significant sequence similarity but which bind different ligands, and it also identified common functional motifs shared by heterogeneous proteins. For example, the pocket similarity was higher for kinase pairs known to share ligand binding preferences than for kinase pairs with divergent ligand specificity. At the other end of the scale, among highly diverse ATP-binding proteins, functional motifs were placed into an accurate taxonomy using the PSIM metric. Sensible segregation was seen between proteins with divergent functions (e.g. serine threonine kinases and tRNA synthetases) as well as among functionally distinct variants within related classes. Extensive comparison to other structural similarity algorithms was presented (e.g. SiteBase,[7] spherical harmonics,[15] sup-CK,[8] PocketMatch,[6] and SMAP[22]), illustrating the complementary of the PSIM local surface similarity approach.

Figure 4.1 illustrates the PSIM computation, showing the optimal alignment of *E. coli* lysyl-tRNA synthetase and human glycyl-tRNA synthetase, proteins with 30%

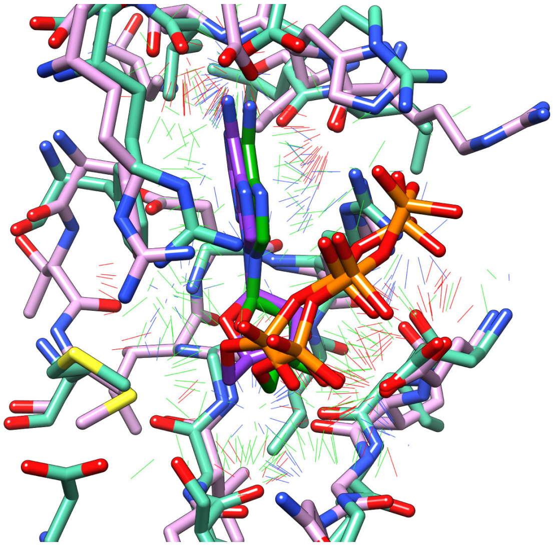


Figure 4.1: Aligned binding sites of two tRNA synthetases. E. coli lysyl tRNA synthetase (PDB code 1E24) is shown in pink, with *H. sapien* glycyl-tRNA synthetase (2ZT7) in teal. The binding site surfaces were aligned using PSIM, with the bound ATP molecules (purple and green) defining the binding site scopes. The thin sticks highlight regions of positive, negative, and steric similarity between the two proteins (blue, red, and green respectively).

sequence identity. The method computes a local comparison of surface shape and electrostatic properties from the vantage point of a putative ligand, irrespective of residue identities or protein backbone correspondence. The two proteins in Figure 4.1 both possess an ATP-binding motif that is characteristic of type II tRNA synthetases. The PSIM taxonomy distinguished this motif from the related, but functionally distinct, type I tRNA synthetase motif. Even in cases where local sequence similarity was non-existent, as observed in some functionally related binding-site pairs from organisms in different kingdoms, PSIM's local similarity computation produced the correct groupings and structural alignments. The ability to make coarse-grained classifications as well as fine-grained distinctions using only binding site similarity suggested the possibility to use PSIM for functional protein annotation.

In the absence of obvious sequence similarity or biochemical experiments, accurate functional annotation requires the ability to directly relate two proteins based on their physical characteristics. PSIM addresses this challenge by evaluating the similarity of binding site surfaces between proteins of interest. The original PSIM algorithm required either a bound ligand or a manually identified location to define a binding site. PSIM has been augmented with an automatic binding site detection feature, removing the need for bound ligands or manual preparation. Screening for functional annotation requires an efficient approach for querying against tens of thousands of protein structures. To this end, the PSIM algorithm has been improved, resulting in a 100-fold speed increase while maintaining performance accuracy. Screening against large libraries also requires the ability to calculate the significance of results given that a large set of purely random comparisons may yield some apparently high scores. A new empirical framework for attributing p-values to protein pair-wise similarity scores provides statistical confidence in PSIM annotations.

The binding site detection methodology was validated on a set of 304 apo/holo crystal structure pairs from the LigASite database.[61] Valid binding sites on apo structures were identified and accurately matched to corresponding sites on holo structures. Site detection performance was equivalent on both holo and apo structures: 91% and 88%, respectively. Further, when querying the holo sites using a detected apo site, we recovered the cognate holo site in 87% of cases. Additionally, similarity scores for apo/holo cognate pairs were statistically separable from scores of random structure pairs (p-value $< 1\times10^{-10}$).

Functional annotation was demonstrated with a screen using a temporally segregated data set. A set of eight proteins whose functions were unknown at the time of deposition were screened against the fraction of the PDB available prior to their definitive biochemical annotation. Using PSIM, the correct function for five of the eight proteins was identified. With respect to functional annotation, PSIM offered complementary information to other established sequence and structural comparison methods (BLASTP,[62] CE,[3] and SMAP[63]). Finally, a screen of twelve uncharacterized proteins against a large database of binding sites resulted in several suggested annotations that merit further investigation.

We believe that the PSIM surface-based approach to functional annotation provides a novel avenue to predicting protein function. PSIM is complementary to sequence-based and backbone-structure-based approaches and may be applied for confirmation or discovery of function. The results provided here demonstrate that the PSIM method may be productively applied on comprehensive repositories of protein binding sites.

## 4.3 Methods and Data

Functional annotation screening required several algorithmic enhancements including: automatic binding site detection, throughput improvements, and a method for determining the statistical significance of a raw similarity score. The new methods along with the data sets utilized for validation and screening are detailed in the following. Data, software, and computational protocols are available by request (see www.jainlab.org for details).

### 4.3.1 Molecular Data Sets

#### 4.3.1.1 Curated LigASite Database

The LigASite Database contains proteins which have been crystallized in both the presence and absence of a bound ligand. When retrieved, the set contained 383 non-redundant pairs of apo/holo structures. Pairs were removed if their ligand did not meet several criteria: a minimum of 7 heavy atoms, at least 3 heavy atoms within 1 angstrom of the protein, and sufficient "buried-ness." The buried-ness of the ligand was measured by taking the ratio of near-ligand protein atoms to the total number of heavy atoms on the ligand. A ratio of 4.0 was used to filter the LigASite database. These various criteria were set such that the ligands present in a standard small-molecule docking data set would be preserved (the Astex85 Set).[64] The majority of ligands that did not pass the filter were ions and crystallization agents with mostly hydrophobic interactions with protein surfaces. The filtered LigASite set contained 304 apo/holo pairs and 606 ligand binding sites.

#### 4.3.1.2 Temporally Segregated JCSG Query Set

The temporally segregated query set was created to validate PSIM's ability to annotate proteins of unknown function without ligands bound. Proteins were identified which had been structurally determined prior to their biochemically annotation. To accomplish this, we queried the Joint Center for Structural Genomics (JCSG) for structures characterized as "Unknown," "Uncharacterized," or "Hypothetical." These queries resulted in 608 matches representing 605 unique PDB identifiers. This set was then reduced to only those structures without bound ligands, eliminating all of the proteins with obvious ligand binding activity or confirmed binding sites. Manual inspection produced 8 unique proteins which were biochemically annotated subsequent to their crystallization.

#### 4.3.1.3 Un-annotated JCSG Query Set

Annotation of currently uncharacterized proteins was performed on an additional set from the JCSG. The un-annotated JCSG query set was created in a similar manner to the temporally segregated set. The 605 structures characterized as "Unknown," "Uncharacterized," or "Hypothetical" were filtered to remove those that had an annotated PFAM domain or an EC number. The 264 structures which remained were filtered again to keep only structures that were determined using X-ray crystallography. The remaining 248 PDB structures came from a variety of sources: bacteria (213), archea (26), eukaryota (6), viruses/other (3). From this pool, 12 structures from model organisms were selected for screening.

#### 4.3.1.4 Curated RCSB PDB Library

Screening was performed against a library of known binding sites derived from the RCSB PDB.[65] All available PDB structures which possessed a bound ligand were procured. This set was then filtered in an identical manner to the LigASite Database, removing those nominal binding sites with very small ligands or which did not form a clear cavity. In total, 30,999 PDB structures with 63,669 binding sites were present in the final library. At the time of curation, this represented roughly 50% of the crystal structures present in the PDB.

### 4.3.2 Computational Methods

#### 4.3.2.1 Surface Based Binding Site Similarity

The core technology of PSIM is a local surface-based molecular similarity computation, which itself was based upon a similarity method for the comparison of small molecules.[10,56] Figure 4.2 depicts the alignment procedure as applied to lysyl-tRNA synthetase and glycyl-tRNA synthetase. Structures typically begin out of alignment with their binding sites in different locations and orientations. The scope of the comparison must be defined by either a bound ligand, a manual selection, or by an automatic binding site detection routine. In this example, the bound ATP molecules are used to define the binding sites on both structures. Each binding site is tessellated with observation points, illustrated by colored spheres in Figure 4.2A. Each observation point measures the distances to the nearest steric, positively charged and negatively charged surfaces. A transformation is found which moves each observation set to the opposing protein and minimizes the differences in corresponding surface measurements (Figure 4.2B). For example, observation points which are initially near a positively charged surface on lysyl-tRNA synthetase are placed as close as possible to similarly charged surface on glycyl-tRNA synthetase. After optimizing the positions of observation points, the transformation is applied to the proteins, placing their binding sites in alignment, as in Figure 4.2C. The similar regions of the protein surface are then depicted using the colored sticks, as in Figure 4.2D. The two synthetases shown have extremely similar binding sites, as indicated by the numerous colored sticks, despite having only  30% sequence identity. The binding site surfaces similarity is amalgamated into a single score ranging from 0 to 1 with this example producing 0.66, representing a high level of similarity. Further details about the computation of protein site similarity scores can be found in our earlier work.[56]

#### 4.3.2.2 Improvements to Surflex-PSIM

The original PSIM algorithm,[56] while accurate, left several opportunities for enhancements to aid in screening applications. While the original computation optimized the alignment of proteins by moving protein atoms, the new method moves the observation points instead. This significantly reduces the number of calculations required, because the average protein has on the order of 10,000 atoms in contrast to the average 400 points in an observation set. Using observation points instead of protein atoms also allows for the geometric hashing of protein coordinates, crucial for the quick lookup of nearby surfaces. In total, these enhancements reduce execution time from 60 seconds per comparison to 0.5 seconds. The results of the new method correlated nearly perfectly with those produced by the previous version (Pearson's r = 0.99).

The definition of the binding-site region has also been improved to allow for two simultaneous and independent sets of observation points. The original PSIM alignment was largely dependent on a single set of observation points. This occasionally allowed for degenerate solutions when working with two binding sites with signifi-

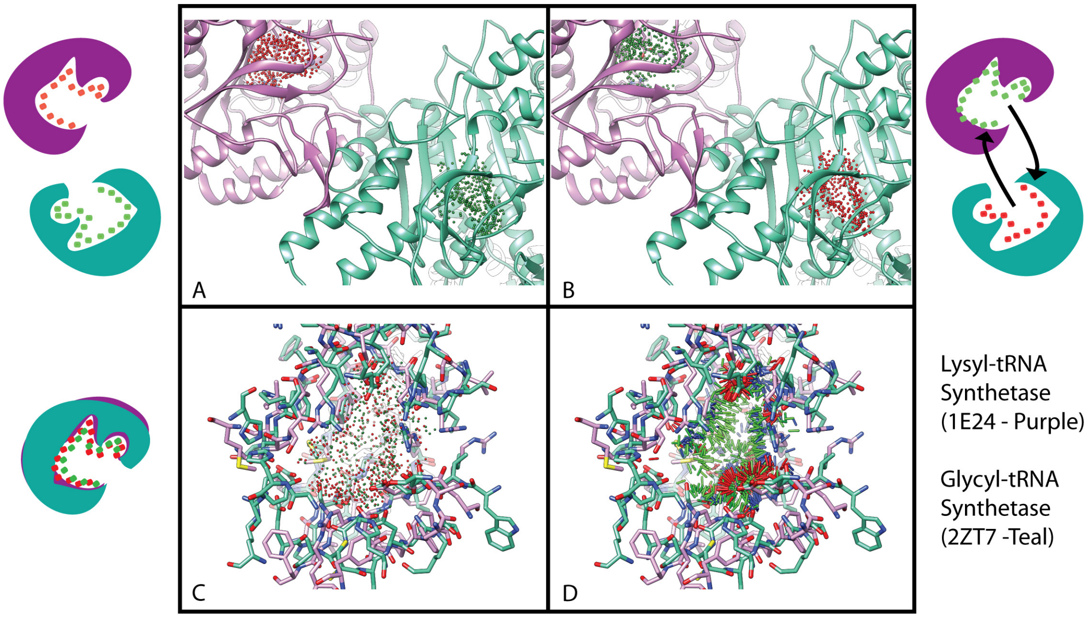


Figure 4.2: The alignment procedure of Surflex-PSIM. *A*: 1E24-LysU in purple and 2ZT7-GlyRS in teal, out of alignment with colored spheres representing the observation points used to measure their binding site surfaces. *B*: A transformation is found which optimally transfers the observations points from their initial locations to similar environments on the other protein. *C*: The alignment used to transform the observation points is applied to the proteins, aligning their binding sites. *D*: Red, blue, and green sticks depict regions of negative, positive, and steric surface similarity, respectively.

cantly different geometry or widely varying size. PSIM now optimizes the placement of both feature sets simultaneously. Using two sets of observation points allows for the geometries of both binding sites to be utilized throughout the optimization process, producing more accurate alignments when comparing divergent sites.

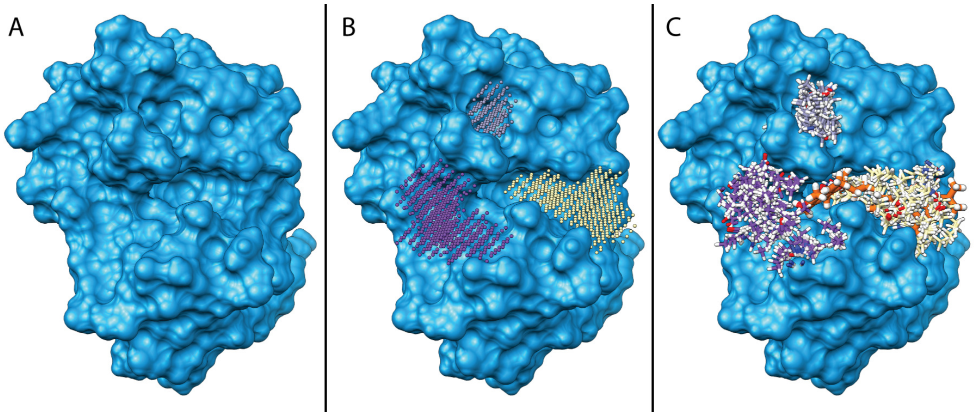


Figure 4.3: PSIM binding site detection. *A*: Bothropstoxin-I (PDB 3CXI) is shown with the bound ligand removed. *B*: Colored spheres indicate discrete regions which passed the concavity filter. *C*: Automatically constructed pseudo-ligands are shown in the concavities identified. Vitamin E, the cognate ligand, is depicted in orange in its bound pose overlapped by the yellow pseudo-ligand (right side).

#### 4.3.2.3 Automatic Binding Site Detection

The PSIM algorithm now includes a procedure for automatically detecting binding sites on protein structures, which is critical for comparing binding sites on structures without bound ligands. An example with Bothropstoxin-I (PDB 3CXI) is shown in Figure 4.3A. The protein is divided into 1 Angstrom voxels marked as either solvent-exposed or internal to the protein. Voxels are ranked by their level of concavity, with concavity measured by counting the number of intersecting paths between pairs of atoms in the region of interest. Paths are assigned a score based on their lengths and the number of internal voxels intersected. Shorter paths that intersect with few internal voxels receive higher scores, and the sum of a voxel's path scores is defined as its concavity. This method parallels previous algorithms in Surflex-Dock[40] and Hammerhead[34,36] for binding site detection and characterization. A fraction of the highest scoring voxels are retained and organized into discrete regions, using a straightforward procedure for connected-components enumeration. In Figure 4.3B, the results of this procedure are shown with each set of colored spheres representing a different concave region.

Each detected connected region is characterized by constructing an optimal pseudo-ligand for the site, again paralleling procedures developed previously for docking.[34,36,40] Each pseudo-ligand is built from three types of molecular fragments: positive polar (N-H), methane ($CH_4$), and negative polar (C=O). All such fragments are placed in candidate binding sites so as to optimize their intermolecular interaction with the protein according to the Surflex-Dock scoring function.[40] Pseudo-ligands with low protein interaction scores or with too few atoms are removed. In Figure 4.3C pseudo-ligands have been built for Bothropstoxin-I; the yellow pseudo-ligand on the right side of the protein occupies the same cavity as the cognate ligand vitamin E (orange).

All parameters for concavity detection, region definition, and binding site filtering were selected so that the ligand binding sites within the Astex85 docking set were identified and retained. Note that the procedure does not identify *only* those sites within the Astex85 set that contain ligands used in docking benchmarking. There are roughly three-fold more nominal binding sites detected than contained bound ligands, and it is not known what proportion of these other sites may have a biological function. The intention for the PSIM algorithm has been to ensure detection of all plausible binding sites rather than to optimize for detection of only those sites where bound ligands have been crystallographically determined.

#### 4.3.2.4 Converting Similarity Scores to P-Values

PSIM raw similarity scores are converted to p-values using an empirical statistical framework and a randomly selected set of unrelated protein binding sites. The random set was created by reducing the PDB library detailed above to proteins with less than 30% sequence identity and by removing proteins which bound the same ligand. This produced 1556 binding sites. A candidate binding site is profiled against this set, producing raw scores, which were observed to be normally distributed. Using a normal distribution to fit the empirical score population, a score resulting from comparison of the candidate site to a protein of interest is converted into a p-value using the analytical integral of the fitted distribution. A typical protein binding site had a mean of 0.48 and standard deviation of 0.026 using PSIM with this random background set. So, a PSIM score of 0.60 would yield a p-value of $1.0\text{e}\times10^{-6}$. The same approach, employing the same background set of protein binding sites, was used to compute p-values for CE[3] and SMAP[22] in the results that follow. The comparison methods CE and SMAP were used as comparators due to their maturity, wide use, and availability.

### 4.3.3 Binding Site Screening

Functional annotation of binding sites was performed in a screen against the previously described library of roughly 60,000 PDB binding sites. Following automatic binding site detection on each query protein, all of the predicted binding sites were compared to the entire binding site library. The resultant scores were transformed into p-values, and those scores passing a threshold were labeled as hits (Z-score > 5.0, equivalent to $p < 2.9\times10^{-7}$). This threshold was chosen such that if no binding site within the library was related to the query, there would be a less than 2.0% chance of returning a nominal hit.

## 4.4 Results and Discussion

The overall aim of this study was to establish the feasibility of a high-throughput protocol utilizing local protein binding site similarity for functional annotation of poorly characterized proteins. Given a high-resolution structure of such a protein from a project such as the JCSG, putative binding site detection yields candidates for screening against a library of well-characterized sites from the PDB. The following details experiments for validation of the binding site detection algorithm, making use of matched apo and holo protein structures. Next, the feasibility of functional annotation is presented, making use of 8 proteins from the JCSG, using a strategy of temporal segregation. Functional annotation evidence for each query protein was derived from protein structures deposited in the PDB *before* the query structure. Finally, the results of a screen for function of 12 proteins that are, as yet, poorly understood, will be presented, providing putative new annotations for these proteins.

### 4.4.1 Binding Site Detection and Apo/Holo Matching

To assess whether the automatic binding site detection and characterization approach would be potentially useful for functional annotation, three aspects were quantified: 1) the detection rate of ligand binding sites in holo protein structures (with ligands removed); 2) the detection rate in corresponding apo structures; and 3) whether the detected sites on apo structures could be correctly matched to their cognate holo binding sites. Performance in the last two assessments was the most important in establishing feasibility for PSIM for the functional annotation application.

#### 4.4.1.1 Detection of Binding Sites on Structures with Ligands Bound

Binding site discovery was performed on 304 holo structures (containing a total of 606 bound ligands), using the parameters derived from initial experiments done on a ligand docking benchmark (see Methods for details). After removal of ligands, the detection algorithm produced a list of putative binding sites, each characterized by a pseudo-ligand that defined its extent. If such a binding site overlapped the volume of

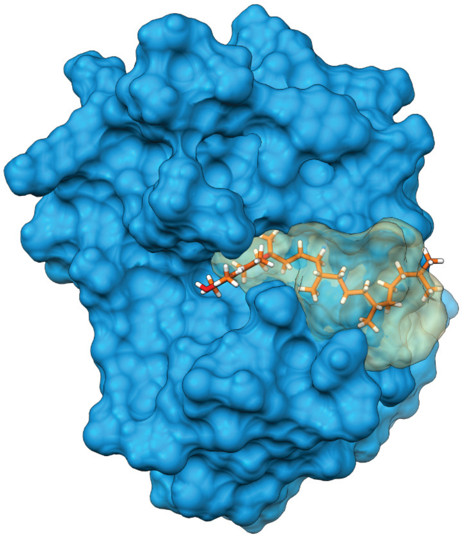


Figure 4.4: Holo protein binding site detection. PDB structure 3CXI (Bothropstoxin-I), shown in blue, is bound to vitamin E, shown in orange sticks. A predicted binding site overlapping vitamin E is shown as a transparent surface.

the known bound ligand by more than 20Å$^3$, the binding site was considered found. Figure 4.4 illustrates a successful case of detection. Overall, the site detection procedure correctly identified the locations of 554 out of 606 ligands, a discovery rate of 91%. In total, slightly more than 4,000 putative binding sites were proposed, yielding a ratio of detected sites to crystallographically observed ligands of about 7:1. It is not possible to know what proportion of the "excess" sites are false positives and which may have an as yet uncharacterized function. However, as mentioned earlier, for a functional annotation application, the critical feature is a high true positive detection rate. Nominal false positive sites contribute to increased computational burdens, but the presence of "decoy" sites on query protein structures can be ameliorated by using conservative cutoffs for defining screening hits.

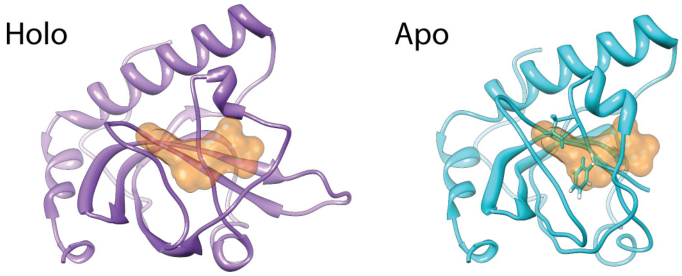


Figure 4.5: Apo binding site collapse. The holo (purple) and apo (teal) structures of riboflavin kinase are depicted in ribbons (PDB codes 1N08 and 1N05, respectively). The area occupied by the ligand (ATP) is depicted in orange. Residues which overlap with the binding site region are depicted as sticks in the apo structure.

#### 4.4.1.2 Detection of Binding Sites on Structures without a Ligand Bound

The detection problem in apo structures is more challenging than for holo structures, in some cases becoming impossible due to binding site collapse. Binding sites which were occluded by molecular motion between apo and holo forms were omitted from this analysis. Overall, 11 protein pairs of the original 304 exhibited very significant conformational movement between apo and holo forms. For example, the pair of riboflavin kinase structures, shown in Figure 4.5, demonstrated a significant side chain motion which completely obscured the ligand binding site. In the holo form, ATP (depicted in orange skin) occupies a tight pocket which, in the apo form, is filled by the rotation of a proline and a phenylalanine (depicted with sticks). Obscured binding sites were identified by aligning the apo and holo structures and determining the amount of protein from the apo structure which overlapped with a ligand from a holo structure. If more than 50% of the ligand volume was occluded, the binding site was considered obscured.

Of the 606 binding sites from the holo detection experiments, 584 were unobscured

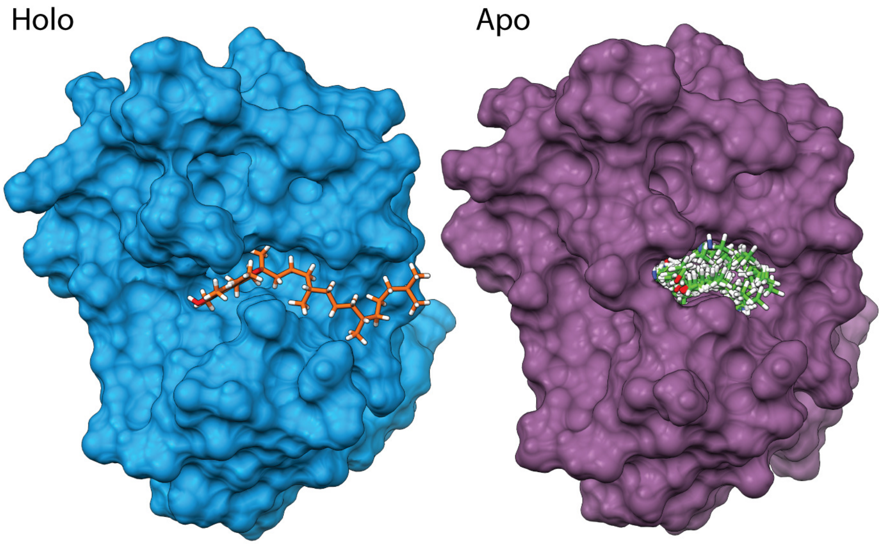


Figure 4.6: Correspondence of apo and holo binding sites. The holo form of Bothropstoxin-I bound to vitamin E is shown (3CXI in blue) along with the apo form (3I3I, in purple). The green molecular fragments on the apo structure depict the binding site detected by PSIM in terms of both location and extent.

in the apo set. Within this set of 584 detectable apo binding sites, the algorithm discovered 514 sites, a discovery rate of 88% (criteria were analogous to the previous experiment). Within the apo set, the ratio of total discovered binding sites to established ones was just slightly lower than in the holo case, roughly 5.5:1. Figure 4.6 presents an example of an accurate recovery of the bothropstoxin-I binding site. Upon alignment of the apo and holo structures of bothropstoxin-I, the holo structure ligand vitamin E overlapped with a detected apo binding site. Note that the overall extent of the apo site, as defined by the green molecular fragments, was smaller than that occupied by vitamin E. Comparison of sites with asymmetrical sizes, which had been a challenge for the original PSIM algorithm, is now addressed with the use of bi-directional site comparison (see Computational Methods). Tolerance to such site variations was critical for efficient recovery of matching sites, as described below.

#### 4.4.1.3 Recovery of Holo Structures using the Apo Structure as a Query

The last aspect of feasibility for protein function annotation requires that a properly detected binding site on an apo protein be correctly matched to a protein binding site with known function. Here, we assessed the likelihood of correctly identifying the cognate partner of each detected apo binding site (514 total) as the top-scoring match in a screen against *all* of the ligand binding sites from the holo structures (606 sites). In 448/514 cases, based on local surface similarity alone, the detected binding sites were correctly matched (an 87% success rate). Further, there was a significant separation in the overall distributions of scores when considering cognate apo/holo pairs and randomly assigned pairs ($p < 1.0\times10^{-10}$ by Wilcoxon rank sum test).

Overall, considering binding site collapse in apo proteins (4% loss), detection performance (88% success), and matching accuracy (87% correct top match), the results suggest a probability of roughly 73% for correctly finding and matching a binding site to its cognate ligand-bound variant. This represents an upper bound on the effectiveness of the functional annotation pipeline. In this scenario, the cognate match will be close to maximally similar to the query under normal conformational variation. In a prospective functional annotation scenario, where no database proteins have significant sequence similarity to the query, it should be expected that the structural similarity of the query binding site to any database member would be lower, thus reducing the chances of identifying a strongly related site.

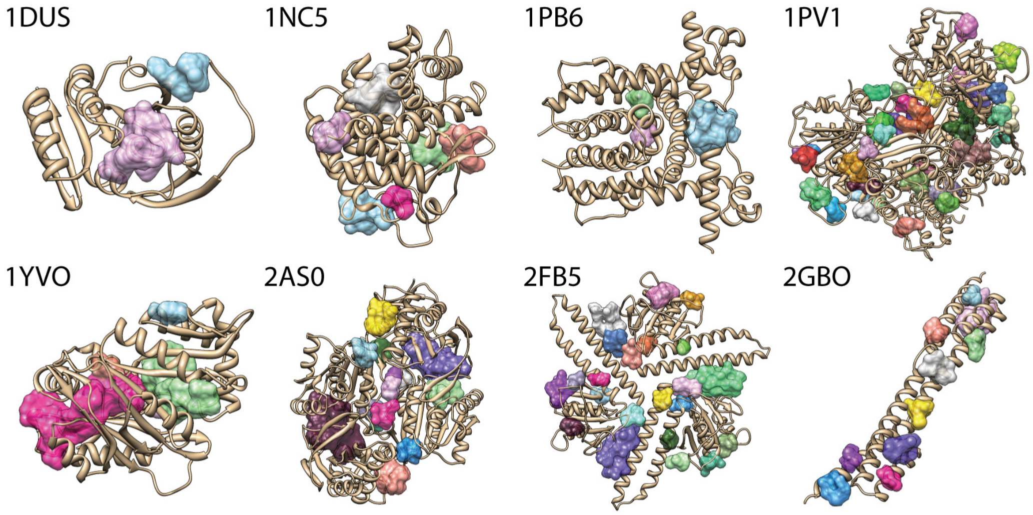


Figure 4.7: JCSG protein structures deposited before they were biochemically annotated. The regions identified as putative binding sites by PSIM are shown in colored surfaces.

### 4.4.2 Analysis of Temporally Segregated JCSG Proteins

Figure 4.7 shows the structures of eight proteins from the JCSG chosen to validate the functional annotation capability of the PSIM algorithm (see Methods and Data for details on the selection protocol). For each protein, site detection was performed using the default parameterization described earlier, and the locations of those putative binding sites are depicted with colored surfaces in Figure 4.7. For each identified protein site, a screen was carried out using PSIM against a library of roughly 60,000 protein binding sites extracted from the PDB. In each case, database proteins were considered whose deposition date was *earlier* than the date the query protein was biochemically characterized. Using this approach, top-scoring matches were returned for 5 of the 8 query structures, providing matches that elucidated biological function. The remaining 3 structures (1NC5, 1PB6, and 2GBO) yielded no matches exceeding the chosen significance threshold. Further analysis showed that 1NC5 and 1PB6 matched variations of the same proteins deposited in the PDB subsequently to the original structures. The following provides details for the five positive cases: 2FB5, 1PV1, 1DUS, 2AS0, and 1YVO.

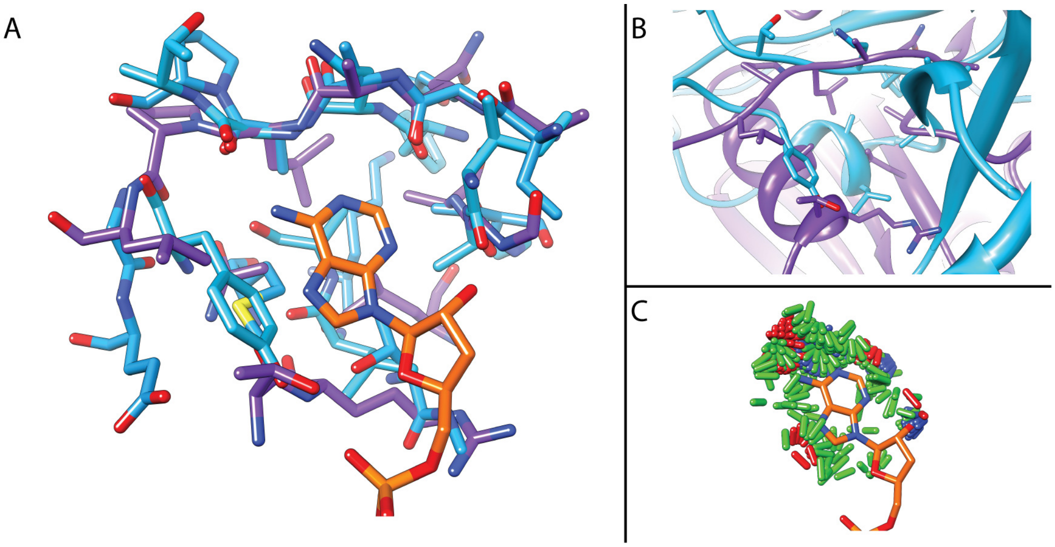


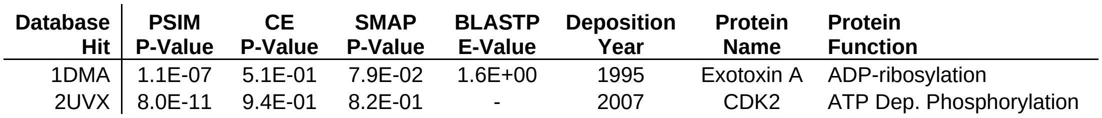
Figure 4.8: Query structure 2FB5. *A*: 2FB5 is shown in purple, along with two screening hits 1DMA (Exotoxin A) and 2UVX (CDK2) in blue. ATP from a later crystallized structure is shown in orange as a guide. *B*: The secondary structure of 2FB5 and 2UVX. *C*: The similarity in the surfaces of 2FB5 and 2UVX is depicted with sticks.

#### 4.4.2.1 2FB5: A Novel Nucleotide Binding Fold

The *Bacillus cearus* protein structure 2FB5 was deposited in 2005 without any knowledge of function. Both the function of the protein and its binding partners remained elusive until homologues of the protein were studied in *Bacillus subtilis* and *Thermotoga maritima* in 2008. These homologues were shown to possess di-adenylate cyclase activity.[66] The ATP co-crystal structures exhibited a novel nucleotide binding fold, which defined a new PFAM motif, designated DisA_N (PF02457, "DisA bacterial checkpoint controller nucleotide-binding").

The PSIM screen for 2FB5 returned two matches to structures that had been deposited *prior* to the 2008 paper (1DMA from 1995, and 2UVX from 2007), correctly identifying the nucleotide binding site (see Table 4.1 for additional details). Figure 4.8 shows these two hits aligned to the 2FB5 binding site. Figure 4.8A highlights a strong similarity in the region corresponding to the hinge of CDK2. This region posses a characteristic alternating set of hydrogen bond donor and acceptor surfaces essential to binding the nucleotide head of ATP. This similarity clearly identifies the presence of a nucleotide binding motif in 2FB5 and is further reinforced by the presence of the same motif in 1DMA (Exotoxin A), where the motif is also used to bind nucleotides.

Table 4.1: Screening results for 2FB5



| Database Hit | PSIM P-Value | CE P-Value | SMAP P-Value | BLASTP E-Value | Deposition Year | Protein Name | Protein Function |
|---|---|---|---|---|---|---|---|
| 1DMA | 1.1E-07 | 5.1E-01 | 7.9E-02 | 1.6E+00 | 1995 | Exotoxin A | ADP-ribosylation |
| 2UVX | 8.0E-11 | 9.4E-01 | 8.2E-01 | - | 2007 | CDK2 | ATP Dep. Phosphorylation |

The SMAP and CE methods yielded weaker p-values than PSIM for these protein comparisons (see Table 4.1), and this is explained by their dependence on congruence of protein backbones. Figure 4.8B shows that the secondary structure of 2UVX is quite distinct from that of 2FB5. There exists almost no secondary structure or atom to atom correspondences in the optimal binding site alignment. However, the local surface similarity depicted in Figure 4.8C shows nearly identical geometry and surface charges in the region around the nucleotide. The green sticks portray a tight cavity formed in the shape of the adenosine, with alternating blue and red sticks indicating the similar "hinge" shared by all 3 structures. For 2FB5, the PSIM approach would have suggested a nucleotide binding function at the time of structure deposition in 2005 through highly significant similarity to 1DMA (which had been deposited in 1997). This may have hastened characterization of this protein.

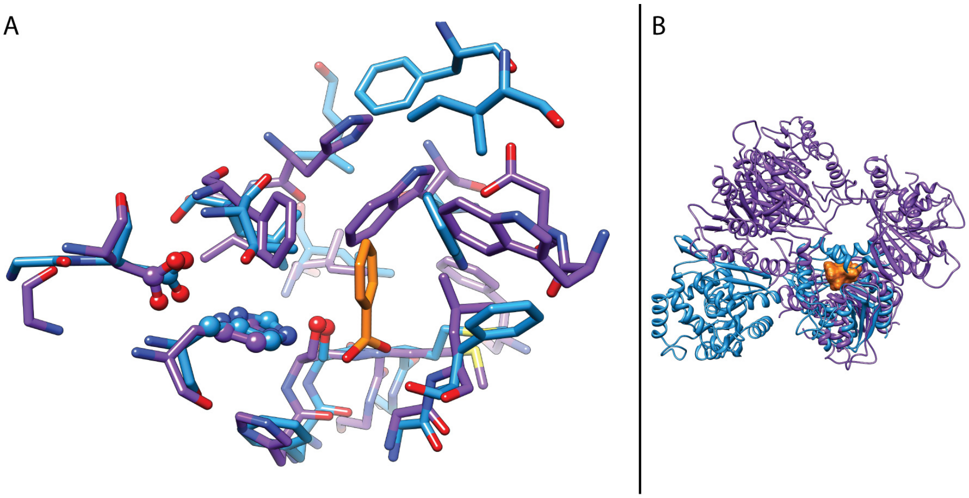


Figure 4.9: Query structure 1PV1. *A*: 1PV1 is shown in purple, aligned with 1A8U (chloroperoxidase T), in blue. The ligand of 1A8U (benzoic acid) is shown in orange, and the catalytic triad (Asp-Ser-His) is depicted in balls and sticks. *B*: The secondary structures of 1PV1 and 1A8U are depicted with the binding site occupied by an orange surface.

#### 4.4.2.2 1PV1: Hypothetical Esterase

The *S. cerevisiae* protein structure 1PV1 was deposited in 2003, having been hypothesized to have esterase activity based on 40% sequence similarity with the known human esterase hEstD. Biochemical confirmation as a carboxylesterase was published in 2008.[67] The PSIM database screen of 1PV1 returned four esterase binding site matches deposited prior to the 2008 publication: chloroperoxidase T (PDB 1A8U, p = $7.1\times10^{-12}$), a still unnamed *Thermotoga maritima* esterase (PDB 3DOI, p = $4.4\times10^{-11}$), butyrylcholinesterase (PDB 1XLU, p = $4.8\times10^{-10}$), and valacyclovir hydrolase (PDB 2OCI, p = $3.7\times10^{-8}$).

The esterase match 1A8U (chloroperoxidase T) was deposited in the PDB in 1998, well before deposition of 1PV1 and thus could have been used to help guide annotation of this protein. Unlike an inference based upon sequence identity, the alignment shown in Figure 4.9A provides additional confidence by showing that the key functional groups required for esterase activity are present and in the correct physical locations in 1PV1. The catalytic triad (lower left, Asp-Ser-His residues) is found in a nearly identical conformation in both 1PV1 and 1A8U. Support for esterase activity also derives from homology and common activity of chloroperoxidase T and bacterial esterases, information available at the time a match could be made for 1PV1.[68] The secondary structures of 1A8U and 1PV1 (Figure 4.9B) show relatively little local similarity, again explaining weak p-values from backbone and sequence based methods (CE: $1.9\times10^{-2}$ SMAP: $2.0\times10^{-2}$ and BLASTP: $1.6\times10^{-1}$).

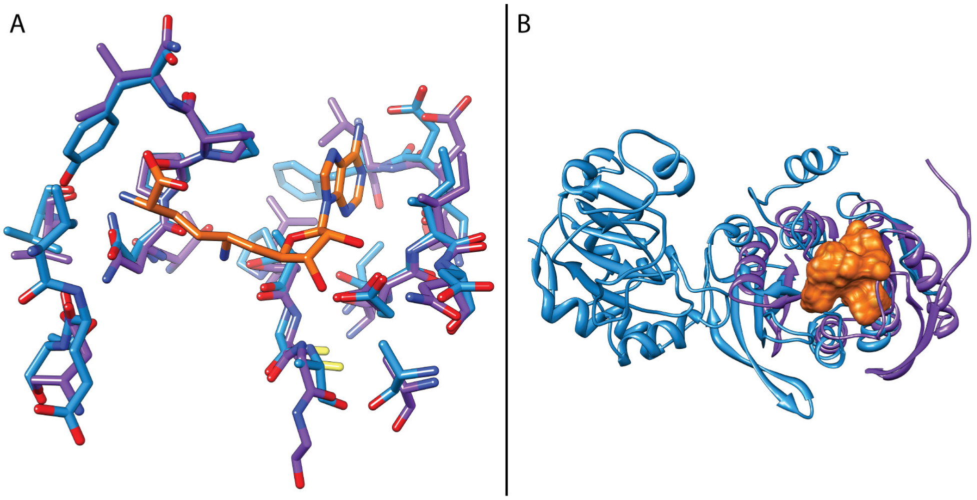


Figure 4.10: Query structure 1DUS. *A*: 1DUS is shown in purple aligned with 1AQJ (DNA-methyltransferase TaqI) in blue. The ligand of 1AQJ (sinefungin) is shown in orange sticks. *B*: The secondary structures of 1DUS and 1AQJ are shown with their binding sites identified (orange surface).

#### 4.4.2.3 1DUS: Subtleties with Structural Comparisons

The *Methanococcus jannaschii* protein structure 1DUS was deposited into the PDB in 2000, and had no sequence similarity to any annotated proteins. Structural similarity methods available at the time suggested similarity to a diverse group of methyltransferase proteins. Further testing and analysis ultimately lead to 1DUS's annotation as a DNA methyltransferase.[69]

Table 4.2: Screening results for 1DUS organized by PSIM p-value

| Database Hit | PSIM P-Value | CE P-Value | SMAP P-Value | BLASTP E-Value | Deposition Year | Protein Name | Protein Function |
|---|---|---|---|---|---|---|---|
| 1G55 | 4.0E-11 | 1.3E-03 | 1.6E-04 | 2.6E-02 | 2000 | DNMT2 | DNA Methyltransferase |
| 1QAO | 6.3E-11 | 8.0E-04 | 1.9E-05 | 2.0E-05 | 1999 | ERMC | RNA Methyltransferase |
| 1QAQ | 1.0E-10 | 8.0E-04 | 2.5E-06 | 2.0E-05 | 1999 | ERMC | RNA Methyltransferase |
| 1AQJ | 9.2E-10 | 1.3E-03 | 7.2E-06 | 1.3E+00 | 1996 | TAQI | RNA Methyltransferase |
| 2HMY | 1.0E-09 | 3.1E-03 | 5.7E-03 | 2.5E+00 | 1999 | HHAI | DNA Methyltransferase |
| 1QAN | 1.1E-09 | 3.1E-03 | 3.0E-05 | 2.0E-05 | 1999 | ERMC | RNA Methyltransferase |
| 1F3L | 5.8E-08 | 3.1E-04 | 9.3E-04 | 8.0E-05 | 2000 | PRMT3 | ARG Methyltransferase |
| 1EG2 | 9.2E-08 | 1.7E-02 | 8.8E-04 | - | 2000 | RSRI | DNA Methyltransferase |
| 1FP1 | 2.6E-07 | 2.0E-03 | 1.1E-02 | 2.6E-01 | 2000 | ChOMT | Chalcone O-methyltransferase |

PSIM returned nine matches that were deposited prior to the deposition of 1DUS. The results are shown in Table 4.2, clearly indicating that this binding site is a methyltransferase. All of the most significant matches were DNA and RNA methyltransferases. These were correctly ranked above proteins such as PRMT3 and ChOMT. The results for CE yielded weaker p-values and also placed PRMT3, an arginine methyltransferase, among the top three hits. For this case, the SMAP method, while producing p-values less extreme than for PSIM, correctly ranked the nucleotide methyltranferases above the others.

The oldest structure which could have provided the correct annotation for 1DUS is 1AQJ, deposited 4 years prior to the deposition of 1DUS. The alignment of 1AQJ and 1DUS (Figure 4.10A) shows a clear correspondence between residues of the uncharacterized protein and the confirmed DNA methyltransferase. Here, in contrast to the previous examples, both the local similarity at the surface and atomic level *and* the local backbone congruence (see Figure 4.10B) are high. Interestingly, though these features are apparent to all of the tested structural methods, BLASTP was unable to detect the similarity between 1DUS and 1AQJ (E-value, 1.3).

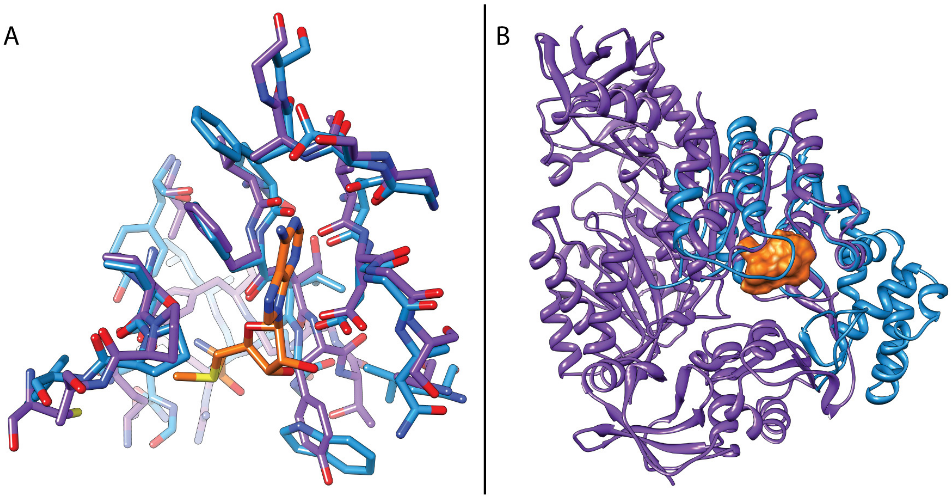


Figure 4.11: Query structure 1AS0. *A*: 2AS0 is shown in purple aligned to 2DPM (adenine-specific methyltransferase DpnII) in blue. *B*: The secondary structures of 2AS0 and 2DPM are shown with the binding sites identified (orange surface).

#### 4.4.2.4 2AS0: Strong Confirmation of Activity

The *Pyrococcus horikoshii* protein structure 2AS0 was deposited in the PDB in 2005. Structure and sequence based evidence suggested its function as a hypothetical RNA methyltransferase.[70] The hypothesis was confirmed in 2008, demonstrating methyltransferase activity specific to 23S rRNA.[71] The PSIM screen of 2AS0 returned nine hits, all indicating methyltransferase activity. All matches were RNA or DNA specific methyltransferases, strongly indicating nucleotide-specific methlytransferase activity. The earliest match was deposited in 1998 (shown in Figure 4.11A). The nearly identical fold structure depicted in Figure 4.11B suggests this to be a particularly easy case for structural comparison methods, and both the CE and SMAP p-values were less than $1.0\times10^{-5}$. Of note, a later variant of the protein represented in the previously mentioned 1DUS example was among the matches (PDB code 2YX1, PSIM p-value: $1.2\times10^{-9}$).

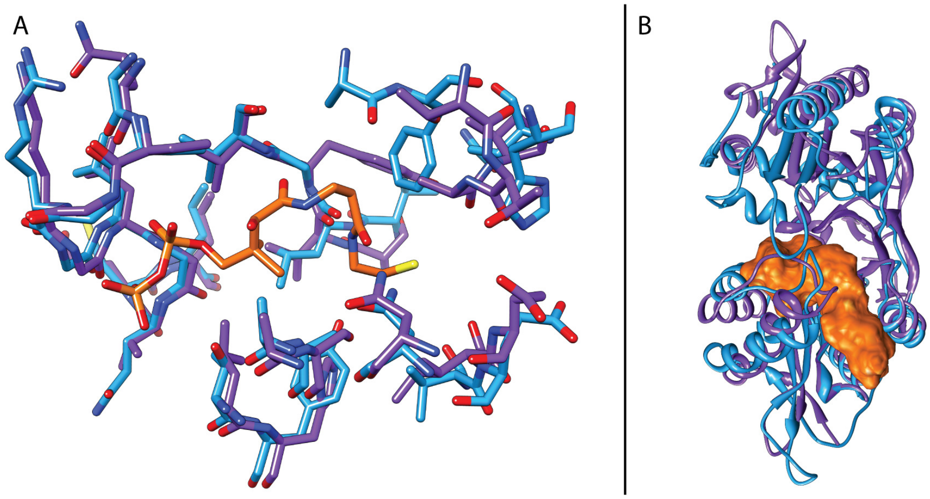


Figure 4.12: Query structure 1YVO. *A*: 1YVO is shown in purple aligned to 2C27 (acetyltransferase, specifically a mycothiol synthase) in blue. *B*: The secondary structures of 1YVO and 2C27 are shown with the binding site identified (orange surface).

#### 4.4.2.5 1YVO: Confirmation of Putative Function

*Pseudomonas aeruginosa* protein structure 1YVO was deposited into the PDB in 2005. The protein shares sequence similarity with a group of genes having mixed functions: some have acetyltransferase activity, and some do not. Because of this conflict, sequence similarity was insufficient for a conclusive assignment of function, but the the acetyltransferase behavior was confirmed biochemically by 2007.[72]

The PSIM screen of 1YVO returned six matches deposited prior to definitive functional annotation of the protein. All matches were to proteins with confirmed acetyltransferase activity. The earliest match 2C27 was deposited in 2005 (see Figure 4.12A) at nearly the same time as 1YVO, but two years prior to the biochemical annotation. The similarity exists within the binding site and also at the scale of secondary structure (see Figure 4.12B). As with the case of the clear catalytic triad motif of 1PV1, the structural similarities here reflect a physicochemical environment in the binding site that is congruent to examples with known acetyltransferase activity. Such similarities were found even in cases where the matched proteins had no homology or sequence identity. For example 2I79, a GNAT-family acetyltransferase, yielded a PSIM p-value of $1.3 \times 10^{-7}$, but its sequence yielded no BLASTP matches against the sequence of 1YVO. In cases like 1YVO, where sequence based annotation is non-definitive, structural methods can be utilized as an orthogonal annotation approach, and they may provide a high degree of confidence, based both on pure statistical arguments and on clear structural correspondences.

### 4.4.3 Library Search of Currently Unknown JCSG Proteins

In the foregoing, the PSIM algorithms for binding site detection and database screening were validated using retrospective experiments. Analysis of performance on the apo/holo detection and matching exercise suggested an upper bound of roughly a 73% chance of correctly identifying a structural match given an un-liganded query protein structure. For five of eight cases (63%), the PSIM method was able to provide correct matches to identify or disambiguate biochemical functions of proteins using structures publicly available *prior* to definitive functional annotation of the query proteins. Given this degree of success, PSIM was also tested on 12 JCSG proteins with no current functional annotation. None of these proteins has either known PFAM domains or sequence similarity to any known protein (see Figure 4.13).

A PSIM database screen analogous to the validation experiment just described re-

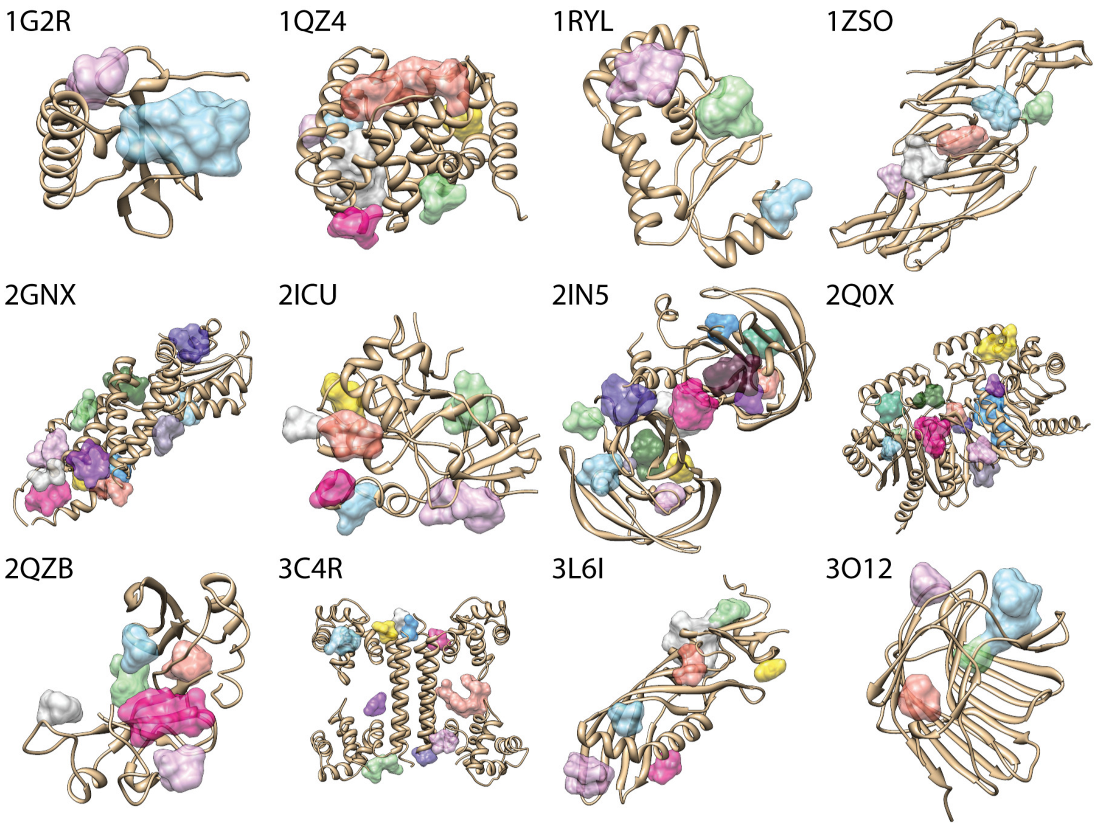


Figure 4.13: Twelve proteins retrieved from the JCSG which are currently uncharacterized. Binding sites predicted by PSIM appear as colored surfaces.

turned numerous statistically significant matches. Table 4.3 summarizes the screening hits obtained for all twelve proteins. The vast majority of nominal hits were matches to shallow surface pockets of proteins within the PDB binding site database or where the bound ligand within the matched site was notably promiscuous (e.g. ethylene glycol, a common crystallization reagent). It is not clear that such sites lack important biological functions, but because their potential functions are not clear, their relevance is difficult to assess. For the query structure 2GNX, which included twelve separate putative binding sites, direct inspection of the matches to identify non-degenerate ones yielded four related hits. Figure 4.14 illustrates the match to the uric acid binding site of 2YZD, with the same coloring scheme as used earlier. The protein is a urate oxidase from *Arthrobacter globiformis*, and it exhibits folds typically found in purine- and pterin-binding enzymes.[73] Other matching structures included three proteins bound to hadacidin (1CG1, 1CG0, and 1KKB, not shown). The three structures are all variants of the same purine biosynthesis enzyme (adenylosuccinate synthetase) from *E. coli*, and hadacidin is an inhibitor of multiple enzymes involved in purine biosynthesis.[74]

Based on the related functions of the matches to proteins found in *Arthrobacter globiformis* and *E. coli*, we suggest that the mouse protein whose structure was deposited as 2GNX (Uniprot ID Q6P1I3) is likely an enzyme involved in some aspect of purine biosynthesis or metabolism. Clearly, this suggestion of potential function is not conclusive. However, it has been included to show that highly significant hits exist for proteins of current interest and to stimulate experimental investigation of the putative functions. All of the data for the database screen of the twelve unknown proteins is included as part of the data archive associated with this paper. Automated methods for detecting deeply buried high-content ligand matches where

Table 4.3: Summary of screening results for 12 unknowns.

| Query Protein | Number of Sites | Hit Count $p < 1.0E-07$ | Hit Count $p < 1.0E-03$ |
|---|---|---|---|
| 2GNX | 12 | 34 | 1652 |
| 1G2R | 2 | 1 | 233 |
| 1QZ4 | 7 | 18 | 860 |
| 1RYL | 3 | 2 | 266 |
| 1ZSO | 5 | 1 | 461 |
| 2ICU | 7 | 7 | 539 |
| 2IN5 | 14 | 4 | 787 |
| 2Q0X | 13 | 12 | 1207 |
| 2QZB | 6 | 15 | 599 |
| 3C4R | 10 | 115 | 1536 |
| 3L6I | 7 | 8 | 796 |
| 3O12 | 4 | 1 | 695 |

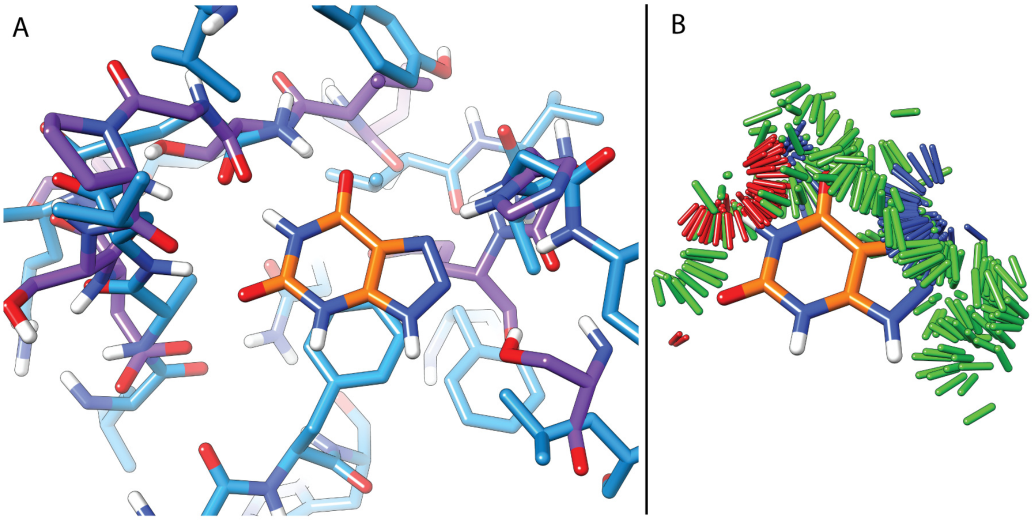


Figure 4.14: Possible function for PDB structure 2GNX. *A*: 2GNX (purple) is shown aligned to 2YZD (blue). The ligand of 2YZD is 8-azaxanthine (orange), and enzyme involved in purine biosynthesis. *B*: The ligand is surrounded by PSIM similarity sticks, depicting a similar steric (green sticks) and polar (red and blue sticks) environment.

multiple matched structures show high mutual similarity at the ligand level will supplant manual inspection of hit lists.

## 4.5 Conclusions

We have presented a fully automatic high-throughput pipeline for protein function annotation through local binding site surface similarity comparisons. The PSIM method is capable of detecting known binding sites in un-liganded protein structures approximately 90% of the time and of matching such sites to their cognate ligand-bound variant with roughly 90% success. Given the occasional occurrence of protein binding site collapse when a binding partner is absent, overall we estimate that 70% success is the upper bound for PSIM to correctly identify a match for a protein site in a large database screen. A query site can be screened against a comprehensive database of ligand binding sites (roughly 60,000 members) in less than one day on a single computing core using standard desktop hardware. Of course, such an application is "embarrassingly parallel" making large screens of many sites against large databases easily prosecuted on computing clusters.

Our initial feasibility studies here on eight JCSG proteins whose functions were definitively assigned subsequent to their structure deposition yielded success in five cases. The most challenging case was one in which a novel nucleotide binding fold was evident only using the local surface similarity method implemented within PSIM. Other structural comparison methods and sequence methods yielded insufficiently strong matches to provide confident functional annotation. In a number of cases, structural comparison methods that relied upon protein backbone similarities yielded solid matches. However, the PSIM approach offered generally stronger evidence in a statistical sense, and it certainly provided an orthogonal means by which to rank the matching protein binding sites. Currently the Protein Structure Initiative[75] contains at least 264 proteins which contain a PFAM domain of "unknown function" or no PFAM domain at all, and these present an opportunity to apply novel similarity metrics for annotation. Our hope is that investigators will be able to make use of the PSIM method along with the PDB ligand binding site database developed for this work in order to aid in protein characterization.

A related area of inquiry involves studies of proteins in dynamic simulation environments, where groupings of related structures may be used for improvements in sampling and for elucidation of transient binding sites.[76] The PSIM approach may offer a different picture from that seen through atomistic clusterings of related protein variants. Apart from questions involving protein function, the PSIM method has clear applications in molecular docking, both for protein alignment and conformational variant selection.[77] This is an area in which many aspects of protein binding site curation and analysis can benefit from fine-grained local similarity computations. For example, it should be the case that clever selection of conformational variants for a particular binding site that *cover* the space of known variations offer a close to optimal choice of protein structures for multi-structure docking.[27]

# References

[1] M. Elsliger, A. M. Deacon, A. Godzik, S. A. Lesley, J. Wooley, K. Wüthrich, and I. A. Wilson. The JCSG high-throughput structural biology pipeline. *Acta Crystallogr., Sect. F*, 66(10):1137–1142, Oct 2010.

[2] L. Holm and P. Rosenström. Dali server: conservation mapping in 3d. *Nucleic Acids Res.*, 38(suppl 2):W545–W549, 2010.

[3] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng.*, 11(9):739–747, 1998.

[4] S. Wu, M. P. Liang, and R. B. Altman. The seqfeature library of 3d functional site models: comparison to existing methods and applications to protein function annotation. *Genome Biol.*, 9(1):R8, 2008.

[5] Nathanaël Weill and Didier Rognan. Alignment-free ultra-high-throughput comparison of druggable protein- ligand binding sites. *J. Chem. Inf. Model.*, 50(1):123–135, 2010.

[6] K. Yeturu and N. Chandra. Pocketmatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinformatics*, 9(1):543, 2008.

[7] S. L. Kinnings and R. M. Jackson. Binding site similarity analysis for the functional classification of the protein kinase family. *J. Chem. Inf. Model.*, 49(2):318–329, 2009.

[8] B. Hoffmann, M. Zaslavskiy, J. Vert, and V. Stoven. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3d: application to ligand prediction. *BMC Bioinformatics*, 11(1):99, 2010.

[9] J. A. Gerlt and P. C. Babbitt. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem.*, 70(1):209–246, 2001.

[10] A. N. Jain. Morphological similarity: A 3d molecular similarity method correlated with protein-ligand recognition. *J. Comput. Aided Mol. Des.*, 14(2):199–213, 2000.

[11] A. N. Jain. Ligand-based structural hypotheses for virtual screening. *J. Med. Chem.*, 47(4):947–61, 2004.

[12] A. N. Jain, N. L. Harris, and J. Y. Park. Quantitative binding site model generation: Compass applied to multiple chemotypes targeting the 5-HT1a receptor. *J. Med. Chem.*, 38(8):1295–308, 1995.

[13] J. J. Langham, A. E. Cleves, R. Spitzer, D. Kirshner, and A. N. Jain. Physical binding pocket induction for affinity prediction. *J. Med. Chem.*, 52(19):6107–25, 2009.

[14] A. Nicholls, G. B. McGaughey, R. P. Sheridan, A. C. Good, G. Warren, M. Mathieu, S. W. Muchmore, S. P. Brown, J. A. Grant, J. A. Haigh, N. Nevins, A. N. Jain, and B. Kelley. Molecular shape and medicinal chemistry: A perspective. *J. Med. Chem.*, 53(10):3862–86, 2010.

[15] A. Kahraman, R. J. Morris, R. A. Laskowski, and J. M. Thornton. Shape variation in protein binding pockets and their ligands. *J. Mol. Biol.*, 368(1):283, 2007.

[16] M. A. Fabian, W. H. Biggs, D. K. Treiber, C. E. Atteridge, M. D. Azimioara, M. G. Benedetti, T. A. Carter, P. Ciceri, P. T. Edeen, and M. Floyd. A small molecule–kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.*, 23(3):329–336, 2005.

[17] L. Hu, M. L. Benson, R. D. Smith, M. G. Lerner, and H. A. Carlson. Binding moad (mother of all databases). *Proteins: Struct., Funct., Bioinf.*, 60(3):333–340, 2005.

[18] P. W. Rose, B. Beran, C. Bi, W. F. Bluhm, D. Dimitropoulos, D. S. Goodsell, A. Prlić, M. Quesada, G. B. Quinn, and J. D. Westbrook. The rcsb protein data bank: redesigned web site and web services. *Nucleic Acids Res.*, 39(suppl 1):D392–D401, 2011.

[19] G. Caetano-Anollés, H. Kim, and J. Mittenthal. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc. Natl. Acad. Sci. U. S. A.*, 104(22):9358–9363, 2007.

[20] S. K. Hanks, A. M. Quinn, and T. Hunter. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science*, 241(4861):42–52, 1988.

[21] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, and R. Lopez. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.

[22] L. Xie, L. Xie, and P. E. Bourne. A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics*, 25(12):i305–i312, 2009.

[23] L. Holm and C. Sander. Dali/fssp classification of three-dimensional protein folds. *Nucleic Acids Res.*, 25(1):231–234, 1997.

[24] I. N. Shindyalov and P. E. Bourne. An alternative view of protein fold space. *Proteins: Struct., Funct., Bioinf.*, 38(3):247–260, 2000.

[25] I. N. Shindyalov and P. E. Bourne. A database and tools for 3-d protein structure comparison and alignment using the combinatorial extension (ce) algorithm. *Nucleic Acids Res.*, 29(1):228–229, 2001.

[26] B. J. Polacco and P. C. Babbitt. Automated discovery of 3d motifs for protein function annotation. *Bioinformatics*, 22(6):723–730, 2006.

[27] A. N. Jain. Effects of protein conformation in docking: Improved pose prediction through protein pocket adaptation. *J. Comput. Aided Mol. Des.*, 23(6):355–74, 2009.

[28] W. A. Hendrickson. Impact of structures from the protein structure initiative. *Structure*, 15(12):1528–1530, 2007.

[29] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161(2):269–288, 1982.

[30] A. J. Olson and D. S. Goodsell. Automated docking and the search for hiv protease inhibitors. *SAR QSAR Environ. Res.*, 8(3-4):273–85, 1998.

[31] D.S. Goodsell and A.J. Olson. Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct., Funct., and Bioinf.*, 8(3):195–202, 1990.

[32] G. Jones, P. Willett, and R. C. Glen. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput. Aided Mol. Des.*, 9(6):532–49, 1995.

[33] G. Jones, P. Willett, R.C. Glen, A.R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267(3):727–748, 1997.

[34] W. Welch, J. Ruppert, and A. N. Jain. Hammerhead: Fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.*, 3(6):449–62, 1996.

[35] A. N. Jain. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput. Aided Mol. Des.*, 10(5):427–40, 1996.

[36] J. Ruppert, W. Welch, and A. N. Jain. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci.*, 6(3):524–33, 1997.

[37] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, et al. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261(3):470–489, 1996.

[38] M. Rarey, B. Kramer, and T. Lengauer. Multiple automatic base selection: protein–ligand docking based on incremental construction without manual intervention. *J. Comput.-Aided Mol. Des.*, 11(4):369–384, 1997.

[39] C. Bissantz, G. Folkers, and D. Rognan. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.*, 43(25):4759–4767, 2000.

[40] A. N. Jain. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.*, 46(4):499–511, 2003.

[41] E. Perola, W. P. Walters, and P. S. Charifson. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins: Struct., Funct., Bioinf.*, 56(2):235–249, 2004.

[42] G. L. Warren, C. W. Andrews, A. M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, et al. A critical assessment of docking programs and scoring functions. *J. Med. Chem.*, 49(20):5912–5931, 2006.

[43] T. A. Pham and A. N. Jain. Parameter estimation for scoring protein-ligand interactions using negative training data. *J. Med. Chem.*, 49(20):5856–68, 2006.

[44] N. Huang, B.K. Shoichet, and J.J. Irwin. Benchmarking sets for molecular docking. *J. Med. Chem.*, 49(23):6789–6801, 2006.

[45] M.J. Hartshorn, M.L. Verdonk, G. Chessari, S.C. Brewerton, W.T.M. Mooij, P.N. Mortenson, and C.W. Murray. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.*, 50(4):726–741, 2007.

[46] J.J. Sutherland, R.K. Nandigam, J.A. Erickson, and M. Vieth. Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy. *J. Chem. Inf. Model.*, 47(6):2293–2302, 2007.

[47] M. L. Verdonk, P. N. Mortenson, R. J. Hall, M. J. Hartshorn, and C. W. Murray. Protein-ligand docking against non-native protein conformers. *J. Chem. Inf. Model.*, 48(11):2214–2225, 2008.

[48] A. N. Jain and A. Nicholls. Recommendations for evaluation of computational methods. *J. Comput. Aided Mol. Des.*, 22(3-4):133–9, 2008.

[49] A. N. Jain. Bias, reporting, and sharing: Computational evaluations of docking methods. *J. Comput. Aided Mol. Des.*, 22(3-4):201–12, 2008.

[50] G. L. Warren, N. Nevins, and G. B. McGaughey. JCAMD special issue: Spring ACS 2011 docking update. *J. Comput.-Aided Mol. Des.*, 26(1):1–2, 2012.

[51] J. B. Cross, D. C. Thompson, B. K. Rai, J. C. Baber, K. Y. Fan, Y. Hu, and C. Humblet. Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.*, 49(6):1455–1474, 2009.

[52] A. N. Jain. Surflex-dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput. Aided Mol. Des.*, 21(5):281–306, 2007.

[53] A. E. Cleves and A. N. Jain. Robust ligand-based modeling of the biological targets of known drugs. *J. Med. Chem.*, 49(10):2921–38, 2006.

[54] A. E. Cleves and A. N. Jain. Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery. *J. Comput. Aided Mol. Des.*, 22(3-4):147–59, 2008.

[55] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.*, 47(7):1739–49, 2004.

[56] R. Spitzer, A. E. Cleves, and A. N. Jain. Surface-based protein binding pocket similarity. *Proteins: Struct., Funct., Bioinf.*, 79(9):2746–63, 2011.

[57] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant. Pubchem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, 37(Web Server issue):W623–33, 2009.

[58] X. Chen, Y. Lin, M. Liu, and M. K. Gilson. The binding database: Data management and interface design. *Bioinformatics*, 18(1):130–9, 2002.

[59] E. Perkins, D. Sun, A. Nguyen, S. Tulac, M. Francesco, H. Tavana, H. Nguyen, S. Tugendreich, P. Barthmaier, J. Couto, E. Yeh, S. Thode, K. Jarnagin, A. N. Jain, D. Morgans, and T. Melese. Novel inhibitors of poly(adp-ribose) polymerase/parp1 and parp2 identified using a cell-based screen in yeast. *Cancer Res.*, 61(10):4175–83, 2001.

[60] P. C. Hawkins, A. G. Skillman, and A. Nicholls. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.*, 50(1):74–82, 2007.

[61] Benoit H. Dessailly, Marc F. Lensink, Christine A. Orengo, and Shoshana J. Wodak. Ligasite: A database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res..*, 36(suppl 1):D667–D673, 2008.

[62] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. Madden. Blast+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009.

[63] L. Xie and P. E. Bourne. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics*, 8(Suppl 4):S9, 2007.

[64] M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. M. Mooij, P. N. Mortenson, and C. W. Murray. Diverse, high-quality test set for the validation of proteinligand docking performance. *J. Med. Chem.*, 50(4):726–741, 2007. PMID: 17300160.

[65] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res..*, 28(1):235–242, 2000.

[66] G. Witte, S. Sophia Hartung, K. Bttner, and K. Hopfner. Structural biochemistry of a bacterial checkpoint protein reveals diadenylate cyclase activity regulated by dna recombination intermediates. *Mol. Cell*, 30(2):167 – 178, 2008.

[67] P. M. Legler, D. Kumaran, S. Swaminathan, F. W. Studier, and C. B. Millard. Structural characterization and reversal of the natural organophosphate resistance of a d-type esterase, saccharomyces cerevisiae s-formylglutathione hydrolase. *Biochemistry*, 47(36):9592–9601, 2008. PMID: 18707125.

[68] I. Pelletier and J. Altenbuchner. A bacterial esterase is homologous with nonhaem haloperoxidases and displays brominating activity. *Microbiology (Reading, Engl.)*, 141 ( Pt 2):459–468, Feb 1995.

[69] L. Huang, L. Hung, M. Odell, H. Yokota, R. Kim, and S. Kim. Structure-based experimental confirmation of biochemical function to a methyltransferase, mj0882, from hyperthermophile methanococcus jannaschii. *J. Struct. Funct. Genomics*, 2:121–127, 2002.

[70] M. Hartshorn, M. Verdonk, G. Chessari, S. Brewerton, W. Mooij, P. Mortenson, and C. Murray. Diverse, high-quality test set for the validation of proteinligand docking performance. *J. Med. Chem.*, 50(4):726–741, 2007. PMID: 17300160.

[71] E. Purta, M. OConnor, J. M. Bujnicki, and S. Douthwaite. Yccw is the m5c methyltransferase specific for 23s rrna nucleotide 1962. *J. Mol. Biol.*, 383(3):641 – 651, 2008.

[72] S. P. Brown and S. W. Muchmore. Rapid estimation of relative protein-ligand binding affinities using a high-throughput version of mm-pbsa. *J. Chem. Inf. Model.*, 47(4):1493–503, 2007.

[73] E. C. M. Juan, M. M. Hoque, S. Shimizu, M. T. Hossain, T. Yamamoto, S. Imamura, K. Suzuki, M. Tsunoda, H. Amano, and T. Sekiguchi. Structures of arthrobacter globiformis urate oxidase-ligand complexes. *Acta Crystallogr., Sect. B: Struct. Sci.*, 64(8):815–822, 2008.

[74] H. T. Shigeura and C. N. Gordon. Hadacidin, a new inhibitor of purine biosynthesis. *J. Biol. Chem.*, 237(6):1932–1936, 1962.

[75] L. K. Gifford, L. G. Carter, M. J. Gabanyi, H. M. Berman, and P. D. Adams. The protein structure initiative structural biology knowledgebase technology portal: A structural biology web resource. *J. Struct. Funct. Genomics*, 13(2):57–62, 2012.

[76] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande. Progress and challenges in the automated construction of markov state models for full protein systems. *J. Chem. Phys.*, 131:124101, 2009.

[77] R. Spitzer and A. N. Jain. Surflex-dock: Docking benchmarks and real-world application. *J. Comput.-Aided Mol. Des.*, 26(6):687–699, 2012.

**Publishing Agreement**

*It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.*

***Please sign the following statement:***

*I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.*

Author Signature

5/28/2013
Date