

FRAGMENT-BASED PROTEIN ACTIVE SITE ANALYSIS USING MARKOV
RANDOM FIELD COMBINATIONS OF STEREOCHEMICAL
FEATURE-BASED CLASSIFICATIONS

A Dissertation

by

REETAL PAI KARKALA

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2009

Major Subject: Computer Science

FRAGMENT-BASED PROTEIN ACTIVE SITE ANALYSIS USING MARKOV
RANDOM FIELD COMBINATIONS OF STEREOCHEMICAL
FEATURE-BASED CLASSIFICATIONS

A Dissertation

by

REETAL PAI KARKALA

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Thomas Ioerger
Committee Members,	Nancy Amato
	Ricardo Gutierrez-Osuna
	James Sacchetti
Head of Department,	Valerie Taylor

May 2009

Major Subject: Computer Science

ABSTRACT

Fragment-Based Protein Active Site Analysis Using Markov Random Field
Combinations of Stereochemical Feature-Based Classifications. (May 2009)

Reetal Pai Karkala, B.E., National Institute of Engineering, Mysore, India;

M.S., Clemson University

Chair of Advisory Committee: Dr. Thomas Ioerger

Recent improvements in structural genomics efforts have greatly increased the number of hypothetical proteins in the *Protein Data Bank*. Several computational methodologies have been developed to determine the function of these proteins but none of these methods have been able to account successfully for the diversity in the sequence and structural conformations observed in proteins that have the same function. An additional complication is the flexibility in both the protein active site and the ligand.

In this dissertation, novel approaches to deal with both the ligand flexibility and the diversity in stereochemistry have been proposed. The active site analysis problem is formalized as a classification problem in which, for a given test protein, the goal is to predict the class of ligand most likely to bind the active site based on its stereochemical nature and thereby define its function. Traditional methods that have adapted a similar methodology have struggled to account for the flexibility observed in large ligands. Therefore, I propose a novel fragment-based approach to dealing with larger ligands. The advantage of the fragment-based methodology is that considering the protein-ligand interactions in a piecewise manner does not affect the active site patterns, and it also provides for a way to account for the problems associated with flexible ligands.

I also propose two feature-based methodologies to account for the diversity observed in sequences and structural conformations among proteins with the same function. The feature-based methodologies provide detailed descriptions of the active site stereochemistry and are capable of identifying stereochemical patterns within the active site despite the diversity.

Finally, I propose a *Markov Random Field* approach to combine the individual ligand fragment classifications (based on the stereochemical descriptors) into a single multi-fragment ligand class. This probabilistic framework combines the information provided by stereochemical features with the information regarding geometric constraints between ligand fragments to make a final ligand class prediction.

The feature-based fragment identification methodology had an accuracy of 84% across a diverse set of ligand fragments and the *mrf* analysis was able to successfully combine the various ligand fragments (identified by feature-based analysis) into one final ligand based on statistical models of ligand fragment distances. This novel approach to protein active site analysis was additionally tested on 3 proteins with very low sequence and structural similarity to other proteins in the *PDB* (a challenge for traditional methods) and in each of these cases, this approach successfully identified the cognate ligand. This approach addresses the two main issues that affect the accuracy of current automated methodologies in protein function assignment.

To Pradeep,

For all the years of strength

ACKNOWLEDGMENTS

As I stand at the end of this endeavor, I am grateful for the love, support and encouragement of so many people: faculty who have taught me to hone my thought processes and to develop as a researcher, colleagues for the comradarie through some of the long nights in front of the computer, friends who have supported me through this entire process with long heart-to-hearts, hugs and sometimes just their presence and family for the continuous support and love without which I would have been lost.

In particular, I would like to thank my advisor Dr. Thomas Ioerger for his continued mentoring and support. I am forever grateful for his introduction of biochemistry that made it possible for me to better understand the basic research in my area and enabled me to contribute more to the research community. Through my many interactions with him, I learned the value of dotting my i's and crossing my t's and being rigorous in my research methodology.

I would like to thank each of my committee members for their contributions to my growth through the graduate process. Dr. Sacchettini showed me that fields of study did not matter as much as the passion for science and the creative thinking to solve a problem. Dr. Guttierrez gave me a solid foundation and understanding of the basic concepts in pattern recognition. There were many a time when his course notes clarified concepts way after I took the class. Dr. Amato has been a source of encouragement and her straight talk helped get me through many a tough time. Through my interactions with her as a committee member and as an AWICS advocate, she opened avenues of growth for me as a researcher and leader.

I would like to thank Janaki Gooty. Her existence has given me hope and strength since I first got to know her 15 years ago and growing old together has certainly been a bonus.

I would like to thank Dr. Williams, my angel of strength. She was my cheerleader, always believing in my capabilities. She took time out to encourage me and I will always cherish the long walks.

I would like to thank Mummy and Daddy for always being in my corner and for always recognizing me. I would like to thank Nischu for her unwavering love and affection over the years and I apologise for the blue bums now. Had I known the gift I was getting, I would have cherished you much more when you were a baby. I would like to thank Sheetu for all the affection and warmth. She brought family to me when she came to the U.S. and the telephone wires could never dampen the strength of her love. And Dhruv, my baby, your chatter has filled my life since you started talking and I have never felt as much joy as I do in every growth of your mind.

Finally, Pradeep, my rock, there are no words that express my joy at having found you.

TABLE OF CONTENTS

CHAPTER		Page
I	INTRODUCTION	1
	A. Previous Approaches	3
	1. Sequence-Based Approaches	3
	2. Structural Information Combined with Sequence Information	4
	3. Combining Sequence Information with 3D Coordi- nates of Amino Acids	7
	4. Docking	10
	5. Previous Feature-Based Approaches	11
	B. Overview of Dissertation	14
II	FEATURE-BASED DESCRIPTIONS OF DIVERSE ACTIVE SITES	19
	A. Molecular Surface and Active Site Surface Generation . . .	20
	B. Active Site Definitions	21
	C. Feature Descriptions	23
	D. Position-Dependent Features Based on Eigenvectors of the Moment of Inertia Matrix	28
	E. Geometric Similarity Analysis	33
	F. Dimensionality Reduction and Classification	33
	1. Singular Value Decomposition	34
	G. Linear Discriminant Analysis	35
	1. Classifier Based on Kernel Density Estimation	37
	H. Conclusions	38
III	COMBINATION OF INDIVIDUAL FRAGMENTS	40
	A. Analyzing Active Site Pockets for the Multi-Fragment Ligands	40
	B. Markov Random Field Theory	42
	1. Formulating as a Labeling Problem	42
	2. Local Neighborhoods to Evaluate Contextual Information	43
	C. Parameter Estimation	56

CHAPTER		Page
	D. Simulated Annealing Algorithm to Sample Conformational Space of Labels	58
	E. Combining Probabilities from Multiple Models into a Unified Prediction of Most Likely Ligand	60
	F. Conclusions	62
IV	DATABASE CREATION	64
	A. Database Creation and Ligand Classes	64
	B. Conclusions	68
V	RESULTS	70
	A. Statistical Comparisons of Classification Accuracies	71
	B. Both Geometric and Electrostatic Features Are Necessary for Classification	72
	C. Note Regarding Current Database	74
	D. Effect of Dimensionality Reduction Techniques Using a Combination of <i>SVD</i> and <i>LDA</i> Projections	75
	E. Fragment Classification Analysis for Classes with One Example	77
	F. Analysis of Classification	80
	G. Fold Family and Homology Analysis	90
	H. Large Active Site Analysis	97
	I. Combination of Fragments Using Markov Random Field	102
	J. Test Cases	105
	1. <i>DEAD</i> Box Protein: 1qde	106
	2. <i>PriA</i> Protein: 2d7h	115
	3. Hypothetical Protein PA1024: 2gjl	121
	K. Effect of Protonation States	124
	L. Application to Drug Discovery	127
	M. Comparison to Previous Methods of Active Site Analysis	129
	N. Conclusions and Future Work	131
VI	SPECIFICITY NORMALIZATION FOR IDENTIFYING SELECTIVE INHIBITORS IN VIRTUAL SCREENING	135
	A. Previous Approaches	136
	B. Methods	139
	1. Linear Programming Formulation	140
	2. Enzyme Assay for Malate Synthase	141

CHAPTER	Page
C. Results	142
1. <i>Rscore</i> for COX-2	144
2. <i>Rscore</i> for DHFR	146
3. <i>Rscore</i> Results for Malate Synthase and Experi- mental Validation	149
4. Promiscuous Virtual Screen Compounds from a Study of ChemBridge Library	153
D. Discussion	159
E. Conclusions	160
VII CONCLUSIONS AND FUTURE WORK	161
1. Future Work	162
REFERENCES	164
VITA	182

LIST OF TABLES

TABLE		Page
I	The number of rotatable bonds and number of conformers generated by Omega for a subset of the multi-fragment ligands	57
II	A sample of the multi-fragment ligands in the current database . . .	65
III	A sample of the ligand fragment classes in the database and a list of the multi-fragment ligands that contain each of these fragments . .	67
IV	Comparisons of classification accuracy using subsets of localized stereochemical features show that classification accuracy is greatly improved when both geometric and electrostatic features are used to describe the nature of the various active site pockets	74
V	Comparison of using localized stereochemical features for fragment classification	84
VI	Top 10 matches and corresponding probabilities for a <i>nicotinamide</i> fragment	86
VII	Results of using geometric and electrostatic position-dependent features for classification	88
VIII	Comparison of leaving fold family out and leaving out homologous sequences during classification using localized stereochemical features	91
IX	Comparison of leaving fold family out and leaving out homologous sequences during classification using position-dependent stereochemical features	94
X	Site-wide accuracy	99
XI	Statistical models for the distances between various fragments in the larger ligands	102
XII	Results of <i>mrf</i> analysis	104

TABLE		Page
XIII	Results of <i>mrf</i> analysis for <i>1qde</i>	109
XIV	Results of <i>mrf</i> analysis for <i>2d7h</i>	119
XV	Results of <i>mrf</i> analysis for <i>2gjl</i>	123
XVI	Differences in classification accuracy using protonated versus de- protonated versions of the electrostatic feature vector	126
XVII	<i>DOCK</i> scores of known COX-II inhibitors across various receptors .	145
XVIII	<i>Rscore</i> calculation and its comparison to <i>DOCK</i> score. Ranks (shown in parantheses) are given as a percentage relative to the ChemBridge library containing 250,000 compounds. μ is the mean average rank over the decoy sites, δ is the difference between the rank against the target receptor and μ , π is the number of recep- tors with positive scores and ϕ is the number of decoy receptors with docking failures	146
XIX	<i>DOCK</i> scores of known DHFR inhibitors across various receptors . .	149
XX	Comparison of <i>Rscore</i> to <i>DOCK</i> score and consensus score for DHFR in virtual screen against ChemBridge library consisting of 250,000 compounds	149
XXI	<i>DOCK</i> scores of known MS inhibitors across various receptors	151
XXII	Comparison of <i>Rscore</i> to <i>DOCK</i> Score and consensus score for MS in virtual screen against ChemBridge library consisting of 250,000 compounds	152
XXIII	% Inhibition for novel inhibitors identified by <i>Rscore</i> ranking	152

LIST OF FIGURES

FIGURE		Page
1	The large increase in the number of new structures in the Protein Data Bank has greatly increased the number of hypothetical proteins (proteins with unknown function)	2
2	Multiple sequence alignment of all the sequences used to define the profile for proteins in the Zinc-Finger family	5
3	An example 3D template for histidine kinase consisting of relative placements as seen in figure of the four residues threonine, histidine, histidine and glycine	8
4	The algorithm flow for <i>CPASS</i>	9
5	One of the conformations adopted by the larger ligand ATP	15
6	Another conformation adopted by the larger ligand ATP. In comparison to the conformation in Figure 5, this conformation has the phosphate moiety of the ligand farther away from the adenine moiety	16
7	The active site pocket for the ligand <i>Adenine</i> bound to the protein 1A4I using the actual ligand coordinates	22
8	The active site pocket for the ligand <i>Adenine</i> bound to the protein 1A4I using a uniform radius of 5Å	22
9	The variation of the largest eigenvalue of the coordinate covariance matrix for the uniform-radius active sites belonging to six fragment classes	29
10	The variation of the cross-sectional feature at 4Å for the uniform-radius active sites belonging to six fragment classes	30

FIGURE		Page
11	The variation of the singular values obtained from an SVD analysis of the training active sites. It shows the significant variation in singular values and the relative importance of information in each of the transformed axes	36
12	Graphical representation of interaction potential functions defined in this study	52
13	Graphical representation of active site mesh to depict intuitively that it is impossible to place the centers of the two fragments of ligand <i>PLP</i> at mesh points i and j without causing steric conflict between these fragments. At the same time, it is quite possible that the fragments are placed at mesh points i and k . Additionally, given the geometric constraints that exist between the placement of <i>PLP</i> fragments, it is impossible for one of the fragments to be placed at mesh point i and the other to be at mesh point l	54
14	The variation of feature difference with distance from actual fragment center for all the fragment classes in this study	55
15	Distribution of number of examples in various fragment classes with fewer than 10 examples	76
16	A combination of <i>SVD</i> and <i>LDA</i> techniques have enabled the selection of features with the most relevant information regarding active site interaction patterns	78
17	Feature-based classification yields a match very similar to the fragment with the highest Tanimoto similarity for a test fragment from a single fragment class	79
18	Match based on feature-based classification is the same as the one with the highest Tanimoto similarity for a test fragment from a single fragment class	81
19	Comparison of Tanimoto scores for classes with one example	81
20	Analysis of database accuracy for various K values	83

FIGURE	Page
21	The variation of feature difference with distance from actual fragment center for all the fragment classes in this study 99
22	Flowchart showing the various steps involved in the analysis of a test protein 107
23	The active site residues identified by superposing the structure of <i>1qde</i> (shown in white) with the structure of <i>2vso</i> complexed to the ligand <i>AMP</i> 108
24	The dimer structure of <i>2vso</i> shows the dimer plays a role in defining the interaction of <i>ribose</i> with the protein 112
25	The active site for <i>1qde</i> with the <i>AMP</i> structure from <i>2vso</i> 113
26	Classification probability peaks for <i>AMP</i> 114
27	The active site residues interacting with the ligand <i>dCMP</i> 116
28	Molecular surface of <i>2d7h</i> showing <i>dCMP</i> in the active site 116
29	Classification probability peaks for <i>dCMP</i> 118
30	The active site residues interacting with the ligand <i>FMN</i> as identified by [47] 122
31	Molecular surface of <i>2gjl</i> showing the deep cleft housing <i>FMN</i> 122
32	Classification probability peaks for <i>FMN</i> 125
33	Comparison of classification accuracies using the methodologies developed in this study, the feature-based methodology (Gutteridge <i>et al</i>) 134
34	Known COX-2 inhibitors used in this study 147
35	The enrichment curves for COX-II based on the three different scores explored in this study. This graph shows that <i>Rscore</i> significantly increases the enrichment in comparison to both <i>DOCK</i> score as well as the consensus score from <i>Sybyl</i> 147
36	Known DHFR inhibitors used in this study 148

FIGURE		Page
37	The enrichment curves for DHFR based on the three different scores explored in this study. This graph shows that <i>Rscore</i> significantly increases the enrichment in comparison to both <i>DOCK</i> score as well as the consensus score from <i>Sybyl</i>	150
38	The four inhibitors identified by <i>Rscore</i> for malate synthase. These four compounds have novel chemical scaffolds when compared to the previously known inhibitors	153
39	The majority of the top 100 virtually promiscuous compounds have a molecular weight greater than 275	155
40	The majority of the top 100 virtually promiscuous compounds have greater than 7 rotatable bonds	156
41	70% of the top 100 virtually promiscuous compounds are either positively or negatively charged. The overall charge for the remaining compounds is zero but they contain both positively and negatively charged components.	157
42	The distribution of polar desolvation energy for the ChemBridge database and the top 100 virtually promiscuous compounds	157
43	This figure shows some of the promiscuous virtual screen compounds identified by our analysis of the ChemBridge library across 10 diverse receptor sites	158

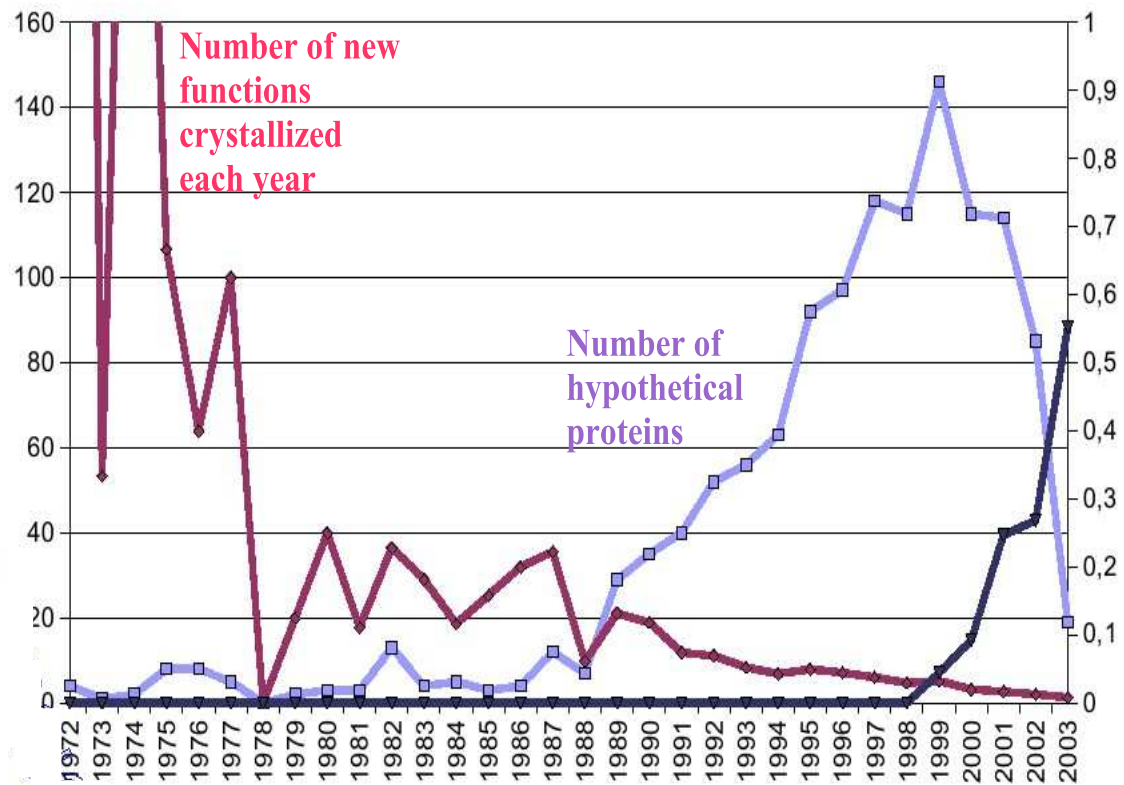
CHAPTER I

INTRODUCTION

The advent of high-throughput structural genomics has essentially changed the process by which biochemists select a protein for study. Previously, a protein was selected based on its functional category for further study and analysis. But high-throughput efforts concentrate on first solving the structures of a large number of diverse proteins and protein function assignment is often based on later studies of the protein structure. Therefore in recent times, functional analysis of 3D structure has become an important problem in structural biology. Figure 1 shows the large increase in the number of proteins with unknown function in the *Protein Data Bank* in recent years [99].

Many computational approaches ranging from sequence-based methods, fold analysis methods to structure-based methods (reviewed in Section A of this chapter) have been proposed for protein functional assignment. However, none of them has been able to capture the diversity in active site geometry and chemistry. In this dissertation, a structure-based approach to functional analysis of proteins based on principles of pattern recognition and machine learning is presented. The function of a protein is based on its interactions with other molecules. Therefore, the identification of a protein's cognate ligand (a ligand that specifically binds to the protein) greatly furthers the knowledge about its function. Here, the functional analysis problem is formulated as a classification problem where the different cognate ligands form the various classes. Given any new protein structure, the aim is to classify it as binding a ligand from one of these classes. Previous approaches to functional analysis have

The journal model is *IEEE Transactions on Automatic Control*.



Pazos and Sternberg; *PNAS* 101(41), 2004

Fig. 1. The large increase in the number of new structures in the Protein Data Bank has greatly increased the number of hypothetical proteins (proteins with unknown function)

largely been unsuccessful due to their inability to address the diversity in active site geometry within a given ligand class. This diversity stems from the conformational flexibility seen in larger ligands (with more than 10 C atoms). In this dissertation, a novel fragment-based approach to dealing with larger ligands is proposed in order to address the problem of flexibility in active site similarity analysis. Larger ligands are broken into smaller fragments consisting of 6-7 C atoms and each of these fragments forms a separate class and multiple ligands can share ligand fragments. This approach necessitates a two-part analysis of any new active site: first, the various possible fragment classes are identified using feature-based approaches and second, these fragment classifications are combined to yield a final large ligand class. Two related but different methodologies for fragment ligand classification and a *Markov Random Field* method for the fragment classification combinations are introduced. In both the methodologies used for fragment classification, active sites are characterized using stereochemical features and relevant features are identified using dimensionality reduction techniques like *Singular Value Decomposition* and *Linear Discriminant Analysis*.

A. Previous Approaches

1. Sequence-Based Approaches

The earliest approaches to functional annotations were based on sequence homology analysis. Sequence-based functional annotation methods are based on the premise that catalytically important residues are conserved in order to preserve function. Proteins that have very high sequence homology ($> 30\%$) or belong to the same structural/fold family tend to have similar physiological ligands (ligand with the maximal binding affinity). Therefore, a straight-forward sequence alignment technique can

be used to identify conserved residues between very closely related proteins. As the sequence similarity decreases, other local patterns need to be identified. *PROSITE* [116] is one such method where motifs of 10-20 amino acids were used to identify catalytically important residues. For all protein sequences that have the same function, a consensus motif is derived based on the active site constituents. Substitutions between residues is allowed at each of the residue positions in the motif based on the BLOSUM [48] or PAM [26] amino-acid substitution matrices. Figure 2 shows the multiple sequence alignment used to define the PROSITE pattern for the Zinc-Finger protein family. *PRINTS* [5] identified that the sequence diversity within a single functional class make it very difficult to explain the active site pattern with a single consensus motif and chose instead to use motif fingerprints or groups of motifs to describe one single functional class. *Pfam* and *PRODOM* databases both use multiple sequence alignments and *Hidden Markov Models* to extract sequence profiles of conserved residues. All of these databases are presently combined as one database resource *InterPro* [3] and this database provides researchers with a resource to identify protein family traits and inherited functional characterization based on sequence motifs.

2. Structural Information Combined with Sequence Information

The problem with purely sequence-based approaches to functional analyses is that very often proteins with highly dissimilar sequences ($< 10\%$ sequence homology) share structural and functional similarities (suggesting a probable common evolutionary origin) necessitating the use of protein 3D structure in functional analyses. Similarly, high sequence homology does not necessarily translate into functional similarity. For example, *NfsA* and *FRP* have over 51% sequence identity but have very different substrate specificities [132]. Therefore, there is a clear need for approaches that

CLUSTAL format alignment

```

AMS2_SCHPO/351-390      -----CQNCGT..I..K..T..AN..WENATY...M.NitLM.LCNACGLIYWTSSRRSMRP-----
AREA_ASPNG/670-723      SSGPTTCINCFT..Q..T..T..PL..WERNPE...G.Q...P.LCNACGLFLKLHGVRPL..SLKTI
AREA_ASPOR/658-711      SNGPTTCINCFT..Q..T..T..PL..WERNPE...G.Q...P.LCNACGLFLKLHGVRPL..SLKTI
AREA_EMENI/667-720      QNGPTTCINCFT..Q..T..T..PL..WERNPE...G.Q...P.LCNACGLFLKLHGVRPL..SLKTI
AREA_GIBFU/688-741      GNAPTTCINCFT..Q..T..T..PL..WERNPE...G.Q...P.LCNACGLFLKLHGVRPL..SLKTI
AREA_PENCH/525-572      -----CTNCFT..Q..T..T..PL..WERNPE...G.Q...P.LCNACGLFLKLHGVRPL..SLKTI
AREA_PENRO/654-707      NAGPTTCINCFT..Q..T..T..PL..WERNPE...G.Q...P.LCNACGLFLKLHGVRPL..SLKTI
ASD4_NEUCR/10-63        ETTQPTCQNCAT..S..T..T..PL..WERDEM...G.Q...V.LCNACGLFLKLHGRPRPI..SLKTI
ASH1_YEAST/493-531      RHTTRVCVSHS..S..D..S..PC..WPSWSprkQ.D...Q.LCNSCGLRYK-----
CGPB_FUSSO/396-429      TSEEYVCTDCGT..L..D..S..PE..WKGPS...GpK...T.LCNACGL-----
DAL80_YEAST/31-78       -----CQNCFT..V..K..T..PL..WERDEH...G.T...V.LCNACGLFLKLHGPRPI..SLKTI
ECM23_YEAST/126-162     NGQPKCATCGD..T..W..T..SQ..WISGPN...G.Nv..E.LCSRCGLIAYR-----
EGL27_CAEL/433-489      GPSGRACHHCYG..A..E..S..KD..WH-HAN...G.L...L.LCTDCRLHYKKYQQLRQI..ANRP
ELT1_CAEL/211-267       STEDRECVNCGV..H..N..T..PL..WERDGS...G.N...Y.LCNACGLYFKMNHARPL..VKPKI
ELT1_CAEL/266-319      KRTGIECVNCRN..N..T..T..TL..WERNGE...G.H...P.VCNACGLYFKLHKVERPI..TMKKI
ELT2_CAEL/231-285      RRQGLVCSNONG..T..N..T..TL..WERNAE...G.D...P.VCNACGLYFKLHHIPRPT..SMKKI
FEP1_SCHPO/9-49         ---GQSCSNCHK..T..T..T..SL..WIRGPD...N.S...L.LCNACGLYQKHKHARPV..-----
FEP1_SCHPO/172-219     -----CQNCAT..T..N..T..PL..WERDES...G.N...P.LCNACGLYKIHGVHRPV..TMKKI
GAF1_SCHPO/629-682      TNPTPTCINCQT..R..T..T..PL..WERSPD...G.Q...P.LCNACGLFMKINGVRPL..SLKTI
GAT10_ARATH/225-261     QYPLRCKMHCEV..T..K..T..PQ..WELGPM...GpK...T.LCNACGLVRYK-----
GAT11_ARATH/193-229     SGGGRRCLHCAT..E..K..T..PQ..WETGPM...GpK...T.LCNACGLVRYK-----
GAT12_ARATH/160-196     QQLRRCSSHCGV..Q..K..T..PQ..WEMGPL...GaK...T.LCNACGLVRFK-----
GAT13_ARATH/215-251     GAEERRCLHCAT..D..K..T..PQ..WETGPM...GpK...T.LCNACGLVRYK-----
GAT14_ARATH/245-281     LQPQRKSSHCGV..Q..K..T..PQ..WAGPM...GaK...T.LCNACGLVRYK-----
GAT15_ARATH/87-123      HSLERRCASCDT..T..S..T..PL..WENGPK...GpK...S.LCNACGLIRFK-----
GAT16_ARATH/37-73       SNEKKSCAICGT..S..K..T..PL..WEGGPA...GpK...S.LCNACGLIRNR-----
GAT17_ARATH/38-74       GDTKRTGVDGCT..I..R..T..PL..WEGGPA...GpK...S.LCNACGLIKSR-----
GAT18_ARATH/163-191     -----NENAT..T..N..T..PM..WIRGPI...GpK...S.LCNACGLIKFR-----

```

Fig. 2. Multiple sequence alignment of all the sequences used to define the profile for proteins in the Zinc-Finger family

combine the sequence information with other structural clues for improved functional prediction/identification.

Initial approaches that defined active site patterns based on a combination of structure and sequence information, used the fold family of the protein as a way to define its structural class [82]. For example, *Rychlewski et.al* [111] studied the *M.genitalium* genome using a combination of a sequence *profile alignment* algorithm and fold-prediction algorithms to tentatively assign function to 80% of that genome. Fold similarity was also used successfully to identify the functional similarity between actin, the ATPase domain of the heat-shock protein, and hexokinase; while these proteins had only a 9% sequence similarity, their structural similarity Z-score was found to be greater than 15 (anything greater than 2 is considered relevant) [16], [39]. There exist many folds like the TIM Barrell that exist in diverse geometries and also combine with different domains to create functionally diverse proteins [90]. Therefore, functional analyses based purely on structural/fold similarity cannot be used to assign biochemical function in an error-free manner. The structural property of solvent-accessibility was used in *ConSeq* [11] in combination with the evolutionary conservation of each residue in order to compute the functional importance of every residue. Structural similarity was also used to identify the functional similarity between *CALB* and *XADL* (*1tca* and *1ede* respectively), two structurally similar hydrolases with dissimilar sequences [40]. The evolutionary trace methodology [76] identified sequence conservation patterns, mapped them onto protein surfaces and compared these mappings to identify functional similarity. This approach was successfully used to identify the functionally important residues in the SH2 and SH3 domains.

3. Combining Sequence Information with 3D Coordinates of Amino Acids

Often, there exists no sequence homology or fold similarity between two functionally similar proteins but the active site residues are conserved as in the case of *PduO-type* *corrinoid adenosyltransferase* from *Lactobacillus reuteri*. A conserved sequence motif was not found in any other existing protein classes and there was no similarity in the overall fold with other known ATP binding proteins [80]. Type I 3-dehydroquinase (DHQase) from *S.typhi* and the type II DHQase from *M. tuberculosis* have totally distinct structures/folds but catalyze the same enzymatic reaction by utilizing completely different mechanisms [45]. Similarly, while the relative positions of active-site components comprising the metal ion Zn^{2+} and *NAD* binding sites were found to be similar in *dehydroquinase synthase* (DHQS) and *alcohol dehydrogenase*, no similarity between the folds of the catalytic domains was found [20]. In order to deal with such cases where proteins from diverse fold families catalyze the same reaction, 3D template methods [61], [8], [9], [70] and [127] were developed. These methods capture the residue patterns between proteins belonging to the same functional class by focusing on the patterns within the active site instead of global structural/fold and sequence similarities. In this methodology, the constellation of residues within the active sites, of all proteins belonging to the same functional class, is described based on the 3D placement of residues as well as the residue identity. Figure 3 shows an example 3D template for *histidine kinase*.

Graph-theoretic approaches, that represent each amino acid in the active site as nodes labeled by the residue identity and relative distances between them as edges in the graph, have been developed [4]. Each such representation of an active site is stored as a graph template. Search patterns, based on the residues in the active sites of new proteins, were also similarly developed and a subgraph isomorphism algorithm

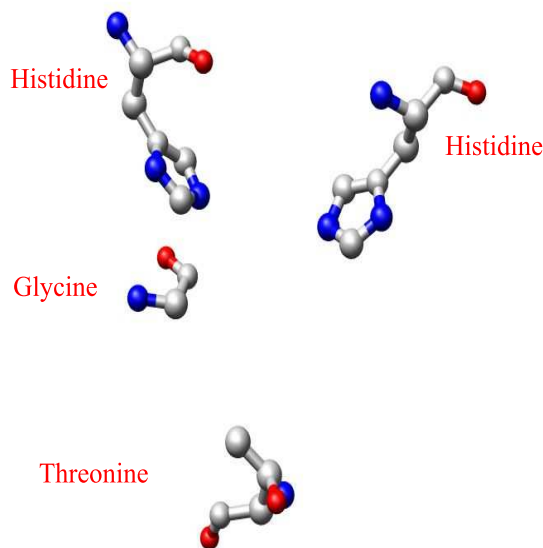


Fig. 3. An example 3D template for histidine kinase consisting of relative placements as seen in figure of the four residues threonine, histidine, histidine and glycine

ASSAM was used to compare between search pattern and the stored templates. The catalytic triads (*SER-ASP-HIS*) found in protease enzyme active sites and the zinc-binding sites in thermolysin were both identified using these 3D template recognition algorithms. The *Catalytic Site Atlas* is the single largest repository (presently containing 18,314 templates) of these 3D templates constructed based on the coordinate data from high resolution structures. The initial 3D template definition of coordinates was very rigid and did not allow for errors that might be present in search patterns due to low-medium resolution data. To account for the variation in interatomic distances due to variations in resolution, *fuzzy functional forms* (*FFFs*) of 3D templates were developed [37]. The *FFFs* seek to relax the structural constraints as much as possible and still maintain the specificity of the active site patterns [83].

The problem with the creation of 3D templates and *FFFs* is the attempt to derive

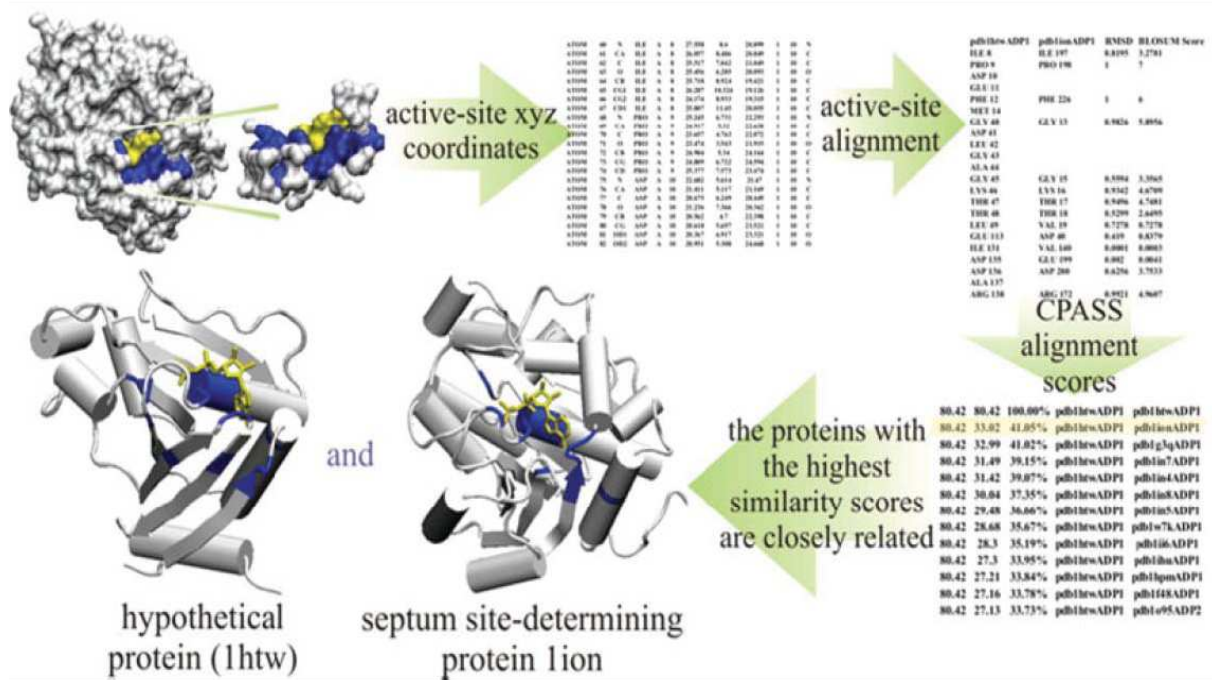


Fig. 4. The algorithm flow for *CPASS*

a generalized template from a diverse set of active sites into a single coherent pattern. This generalization results in a loss of details regarding protein-ligand interactions, since few or no individual contacts might be found common to all members of these protein families. This observation was confirmed by the study on adenine binding motifs [28], [29]. An algorithm *CPASS* [102] was developed to avoid this problem by not creating a consensus template but storing each template representative of a functional class. A new protein is then searched for each of these stored templates. Figure 4 shows the flow of the algorithm for *CPASS*.

There is a large diversity in active sites that extends beyond sequence homology or structural architectures due to the fact that proteins often make do with various residue combinations and structural motifs to effect the same chemical specificity. For

example, the diversity in adenine active sites has been extensively studied and common motifs (a supersecondary β -loop- β that grasps the adenine ring along its faces, base-stacking of the adenine ring with the protein hydrophobic atoms and backbone polar interactions and nonspecific hydrophobic interactions) were found ([69]) and none of these motifs are based on specific amino acid identities and placements. Similarly, the active sites of the DJ-1 superfamily (which consists of kinases involved in the biosynthesis of thiamine [130]) and the active sites of β -carbonic anhydrase from *M.thermoautotrophicum*, *M.tb.* and *B.subtilis* [63] all maintain overall shape and chemical complementarity between active site and ligand despite substantial differences in specific residue identities and placements. Similarly, the overall active site structure of *oxygen-insensitive nitroreductase* (NfsA) from *E. coli* is similar to the NADPH-dependent *flavin reductase* of *V. harveyi*, despite definite difference in the spatial arrangement of residues in the active site [65]. However, this preservation of complimentary interactions between receptor and ligand is not captured by 3D motifs since they rely on rigid motif definitions based on the identities and placements of residues involved in protein-ligand interactions. Generalizing these patterns by allowing fuzzy descriptions of geometry does not increase the accuracy of this approach either. Instead, generalizing only reduces the discriminatory power of a pattern by causing unrelated active sites to look similar, thereby increasing the discrepancies within functional assignments.

4. Docking

Based on previous discussions, it is essential to develop an analysis of active sites that captures the diversity within active sites binding the same ligand and does not rely on specific residue placements and fold analyses. Such an analysis will help to better understand and capture the underlying geometric and chemical interaction

patterns within the active site. According to the *Gibbs* free energy of binding ($\Delta G = \Delta H - T\Delta S$), favorable interactions between ligands and their proteins are determined by the balance of enthalpic and entropic forces acting on the ligand and the protein active site. The free energy of binding equation suggests that high binding affinity only requires sufficient chemical interactions to be accumulated throughout the active site. These favorable interactions do not require specific side-chain position or identity information. A high-affinity protein-ligand complex maximizes favorable chemical interactions and minimizes steric conflicts.

Alternative approaches to functional analysis based on the free energy of binding have been developed, notable amongst them are the docking algorithms like *DOCK* [81] (scoring function evaluates the electrostatic and van der Waals interactions between the protein receptor and ligands), *FlexX* [51] (accounts for flexibility in receptors thereby allowing for the use of apo-structures in functional analysis), *GOLD* [57] (a genetic algorithm to evaluate the interactions between a receptor and ligand), *AUTODOCK* [85] (a lamarckian genetic algorithm) etc. The free energy of binding allows for the active site interaction to be represented in a far more general manner than the residue template approach [114] and [118]. Docking algorithms were developed to attempt to approximate the free energy of binding. Docking algorithms have successfully been used to predict substrates for newly solved structures with low sequence and fold similarity [120], [50]. Unfortunately, this computationally rigorous and time-intensive analysis is not always capable of identifying the correct substrate since accuracy of force fields and scoring functions is still under debate.

5. Previous Feature-Based Approaches

Biochemists have long analyzed active sites by looking at geometric and chemical characteristics/features of the protein-ligand interaction and not restricting this analysis

to exact residue identities or residue placements. For example, *DsbA* is a protein-folding catalyst from the periplasm of *E.coli* that interacts with newly translocated polypeptide substrate and catalyzes the formation of disulfide bonds in secreted proteins. The biochemical analysis of this protein identified three unique features in the protein active site: a groove, a hydrophobic pocket, and a hydrophobic patch, all of whom formed an extensive uncharged surface surrounding the active-site disulfide. Computational approaches mimic this structural analysis with the use of various features. Such an analysis is not only computationally efficient but also abstracts well over diverse active sites that do not preserve sequence identities and placements.

Features allow for broader similarities between active sites to be identified and compared for functional analysis. Authors in [46], [41] and [9] developed high-level features like residue type (charged, polar, hydrophobic etc), solvent accessibility, secondary structure type, conservation etc., to characterize chemical and geometric properties of active sites. These features were then used to train a neural network to identify active sites on a protein surface. None of these features depended on the precise location of residues within the active site. Therefore, they were capable of generalizing over diverse families of proteins. Surface patch analysis also used similar features to successfully characterize protein-protein interactions [58] as well as to distinguish between carbohydrate binding patches and the rest of the surface patches, all obtained from protein-carbohydrate complexes [126]. In all these applications, the features were developed to capture the global characteristics of the active site and therefore could not capture the spatial variations of chemical and geometric properties within an active site. This limits the use of these feature-based methodologies to differentiate between active sites belonging to two different ligands.

FEATURE [6], [7] attempted to characterize local variations in the active site by defining distributions of residue properties in radial shells to capture the differences in

protein microenvironments between protein active sites and non-sites. They successfully used this system to identify common biochemical properties within the serine protease active sites and calcium binding sites. However, their features still relied on specific amino acid identities and therefore did not extend well to cases where high ligand promiscuity allowed little or no sequence conservation between active sites of the same functional class.

Feature vectors composed of 3D moment invariants have been used to capture the geometric shape of protein-protein interaction sites [119]. Similar geometric shapes lead to similar feature vectors, thus enabling the identification of similar binding sites. While, this approach works for protein-protein interaction, its inability to combine chemical information with the shape analysis limits its suitability in recognizing protein-ligand interactions and further, in distinguishing between active sites belonging to two different ligands and therefore two different functional classes.

Computationally intensive procedures like superposition of the active sites [67] or computation of alpha shapes [75] have been used to capture shape similarity between protein-ligand interaction sites. In order to reduce the computational intensity, a recent paper, [60] used spherical harmonic expansion coefficients as descriptors of the active site shape, allowing for shape comparisons without the need for computationally intensive superposition calculations. While successfully differentiating between active sites belonging to ligands with large differences in shape, the authors identified that significant variations exist in the shape of active sites binding the same ligand due to the flexibility observed in larger ligands. This is an important and as yet unaddressed issue with active site shape comparison techniques.

B. Overview of Dissertation

Protein interactions with other molecules change over evolutionary time. In order to preserve essential function, despite mutations, proteins develop multiple interaction patterns with ligands in order to perform the same function. This causes great diversity in active site electrostatics which cannot be captured by simple sequence patterns/templates. Additionally, both the protein as well as the ligand have conformational flexibility, i.e. change in shape so as to better interact with each other. This flexibility in ligands as well as the receptor (protein) have not been successfully modeled by previous computational functional analysis tools. These complications cause the automatic determination of protein function from structure to remain an open problem despite the development of many computational algorithms to analyze protein function. In this dissertation, a machine learning framework to analyze diverse active sites while taking ligand flexibility into account will be proposed (this study does not address receptor flexibility).

The active site analysis problem is formalized as a classification problem where for a given test protein the goal is to predict the class of ligand most likely to bind the active site based on its stereochemical nature and thereby define its function. One formulation of this problem would be to categorize each ligand into its own class, but this model is too simplistic and therefore, the second model would be to group ligands with chemical similarity into a single class collect examples of protein-ligand complexes for each ligand class. Assuming that no cutoffs for sequence homology or structural similarity are enforced while building the database, this approach will capture the diversity within active sites binding the same ligand accurately. But, this model does not take into account the flexibility observed in ligands.

Ligand flexibility causes great variation in active site geometry. As the number

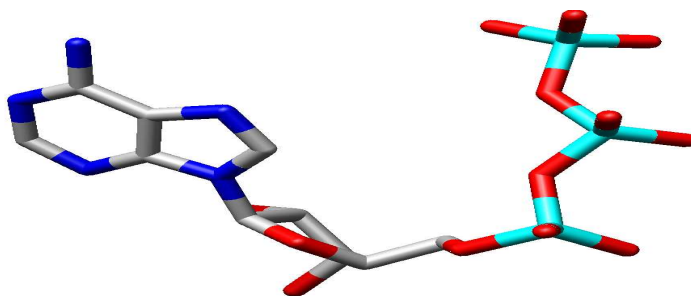


Fig. 5. One of the conformations adopted by the larger ligand ATP

of rotatable bonds in a ligand increases so do its number of possible conformations in a complex. These significant variations in the geometry of active sites binding the same ligand due to the ligand flexibility were studied in a recent paper [60] that used spherical harmonic expansion coefficients as descriptors of the active site shape (allowed for shape comparisons without the need for computationally intensive superposition calculations). Authors in [123] studied the different conformations of *adenine ribose triphosphate* (ATP) bound to various proteins and found an average RMSD deviation of 2.2\AA between conformations. Figures 5 and 6 show an example of the conformational flexibility seen in ATP. In Figure 5 the phosphate moiety is closer to the adenine moiety as compared to Figure 6. This example showcases the need to address the issue of ligand flexibility in any study of protein active sites.

Therefore, a third model to describe protein-ligand interactions is proposed here. Each ligand is divided into fragments (building blocks/subcomponents of ligands, often found in multiple ligands) containing no more than 6-7 C atoms. This limits the number of rotatable bonds and therefore the effects of flexibility on the protein-fragment interaction. Dividing a ligand into fragments does not affect the active site patterns since a strong interaction between a protein and a ligand requires shape

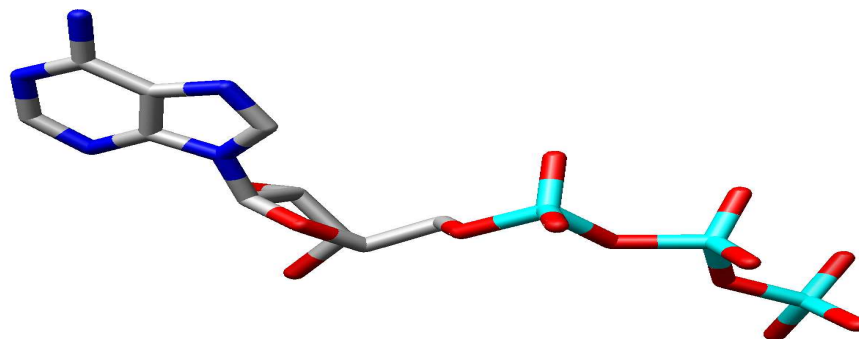


Fig. 6. Another conformation adopted by the larger ligand ATP. In comparison to the conformation in Figure 5, this conformation has the phosphate moiety of the ligand farther away from the adenine moiety

and electrostatic complementarity throughout the active site [104]. Each fragment of the ligand occupies a region of the active site based on local complementary interactions with the underlying protein and interaction patterns for each fragment can be analyzed separately.

Once each ligand has been decomposed into fragments, similar fragments are clustered together based on chemical similarity and each cluster is a class in the classification scheme proposed here. A database of complexes bound to each of the ligand fragments is collected and analyzed for interaction patterns using feature-based methods. Since, the classification algorithm returns a fragment class instead of a ligand class, there are two challenges when analyzing a test protein. The first, is to find the fragments that have high likelihood of binding to the site based on the similarity of stereochemical patterns in the test active site to those in the database and the second is to combine these fragment classifications into the most likely ligand class. The rest of this dissertation will explore these ideas in detail and examine the

efficacy of this approach to automated functional analysis.

In Chapter II, the generation of protein molecular surfaces and active site pocket definitions required by the active site analysis described in this dissertation will be described. The chapter will also describe the methods associated with the description and classification of ligand fragments. Specifically, this chapter will detail two systems of stereochemical features to capture interaction patterns. The first, localized stereochemical features, are an extension of previous feature-based active site descriptions. These features capture local variations in active site geometry and electrostatics instead of just using global descriptors of the active site. These rotation-invariant features allow for a finer-grained analysis of active sites, but, since they are rotation invariant, there is still some information about active site stereochemistry that is lost. The second, position-dependent features are an extension of the localized features designed to capture the variation in stereochemistry specific to active site position. To this end, they are based on canonical representations of the active site. This chapter also details dimensionality reduction techniques used to find the features that contain the most information pertinent to active site description and classification. Finally, this chapter describes the classification algorithm used to identify the fragments most likely to bind an active site.

In Chapter III, a methodology to combine individual fragment classifications into a final ligand classification based on *Markov random field* (MRF) theory is described. This probabilistic framework combines the information provided by stereochemical features with the information regarding geometric constraints between ligand fragments to make a final ligand class prediction.

In Chapter IV, the various ligands as well as ligand fragments used in this study are detailed and the creation of a protein-ligand database is discussed. This database will be used to test the accuracy of the feature-based methodologies for active site

description, test the accuracy of the classification algorithms as well as test the accuracy of the *MRF* model. Chapter V will include the results of all of these analyses and also include the analysis of two test proteins to validate the approach presented in this dissertation.

Finally, in Chapter VI, another novel algorithm, *Rscore*, to improve the efficiency of docking, a previously well-established procedure for functional analysis is described. This chapter will also present results that experimentally validate the *Rscore* algorithm.

CHAPTER II

FEATURE-BASED DESCRIPTIONS OF DIVERSE ACTIVE SITES

In this chapter, the various methodologies involved in the analysis of protein active sites in this study will be introduced. The analysis begins with the definition of a protein molecular surface in Section A and this definition will be then used to define the active site surface as detailed in Section B. This active site surface is used in all future analyses using stereochemical features. As discussed in Chapter I proteins evolve multiple interaction patterns with their cognate ligands. The diversity in these interactions makes it harder to characterize and recognize these interactions. Machine learning techniques especially feature-based methods (described in Section 5 of Chapter I) have been used previously with some success in categorizing various active sites. These previous feature-based methodologies have focused on global features describing the geometric and electrostatic nature of active site surfaces. Unfortunately, these features are unable to capture the diversity in active sites binding the same ligand. Authors in [6] and [7] first considered the use of micro-environments in the description of phosphate binding sites. In this chapter, their methodology is extended to define localized stereochemical features that capture the diversity in the protein-ligand interaction patterns. Section C details the various stereochemical features used in this study. These stereochemical features are rotation-invariant and seek to capture local variations in interaction patterns. Unfortunately, since these features are rotation-invariant they still do not give a detailed description of the active site patterns. Therefore, the localized stereochemical features are extended by using position-dependent features. These position-dependent features are obtained by first using the eigenvectors of the active site pocket moment of inertia matrix to superpose all the active sites into a canonical position. Then these canonical rep-

representations are analyzed to yield a position-dependent (based on 3D coordinates) description of the active site patterns. Section D details the methodology used to generate the position-dependent features.

These feature-based descriptions of the active site are used to classify active sites into the class of their cognate ligand. Since, not all of the developed features contain information relevant to classification, dimensionality reduction techniques need to be employed to identify features that contain information and are relevant to classification. Section 1 describes the classification scheme as well as the dimensionality reduction techniques used in this study.

A. Molecular Surface and Active Site Surface Generation

The first step in the automated analysis of protein active sites is to define the active site surface for each of the proteins in the database. This process begins with the definition of the protein molecular surface. The protein coordinates were used to compute a molecular dot surface similar to the solvent-accessible surface, defined by Richards [105] and later implemented by Connolly [24], using *Calcsurf*, an in-house program. *Calcsurf* simulates the contacts a water molecule (probe sphere of radius 1.4Å) would make with the protein molecule. Considering the radius of the water molecule and the van der Waals radii of protein atoms, a grid representing the dot molecular surface is drawn at a distance equal to the sum of these two radii from the protein molecule. The grid points are spaced 1Å apart allowing a fine-grained representation of the solvent-accessible surface. In the case of proteins where the active site is at the interface of multiple chains, the molecular surface was drawn over all the chains that participate in the active site creation thus allowing for the analysis of such active sites.

B. Active Site Definitions

The final aim of this dissertation is to analyze the function of unliganded proteins. Therefore it is necessary to have a definition of the active site surface that does not depend on the exact coordinates of the native ligand, since this information is unavailable in the case of apo-proteins. Further, the active site could have additional *buffer zones*, *i.e.*, regions of empty space where no ligands bind (as noted by [60]). For a successful application of the methodologies introduced in this dissertation to functional annotation, it is essential that the analyses be robust to the slight noise in the active site patterns introduced due to the inaccuracies in the initial description of the active site. Therefore, the active sites are defined as uniform-radius pockets in order to increase the generality of the active site definition. These uniform active site pockets are created by first choosing a surface vertex closest to the ligand center as the center of the active site. All surface vertices within a chosen radius are then considered to be part of the active site. The choice of the radius depends on the statistical analysis of the fragment pockets in our database (analysis of the average distance of an active site vertex from the center of the ligand). For example, the average distance of an active site vertex from the center of ligands like adenine, citrate, pyridoxal, etc., was found to be 5Å. This definition of the active site as a pocket of a uniform chosen radius, introduces variations in active site shape. Figure 7 shows the active site surface based on the coordinates of the ligand *Adenine* bound to protein 1A4I and Figure 8 shows the active site surface based on a uniform radius for the same protein-ligand complex. A comparison of these two figures shows that there is a loss in specific active site shape information with the use of the uniform radius description. While this introduction of noise into the active site definition makes the problem of active site recognition harder, it increases the utility of this approach to

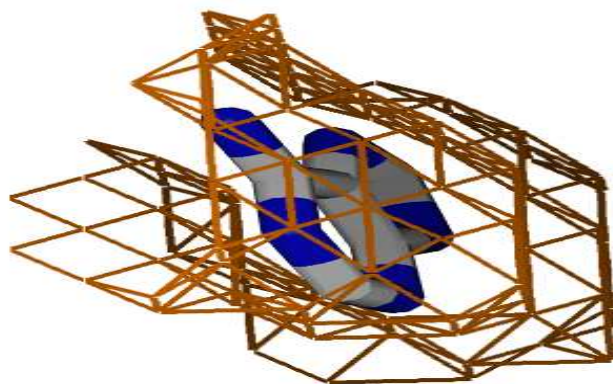


Fig. 7. The active site pocket for the ligand *Adenine* bound to the protein 1A4I using the actual ligand coordinates

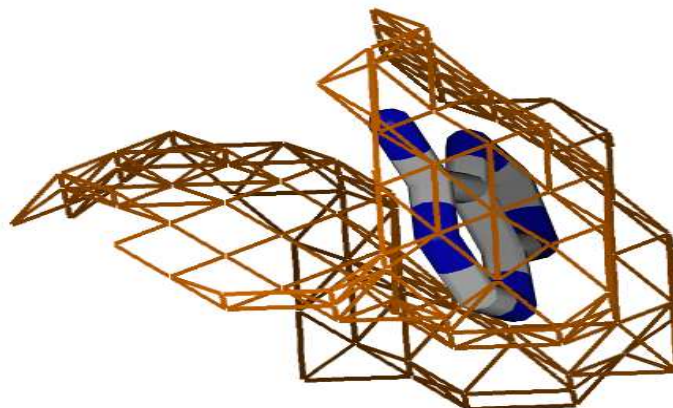


Fig. 8. The active site pocket for the ligand *Adenine* bound to the protein 1A4I using a uniform radius of 5Å

the analysis of active sites in unliganded (apo) proteins where the exact active site shape is rarely known.

C. Feature Descriptions

The global shape features used to describe the uniform-radius active site surface, S , as defined in Section B are as follows:

- The eigenvalues of the coordinate variance-covariance matrix are used to define the spread of the pocket in three dimensions. The eigenvalues λ_1 , λ_2 and λ_3 of the variance-covariance matrix \mathbf{C} are calculated using the following equation:

$$|\mathbf{C} - \lambda\mathbf{I}| = 0 \quad (2.1)$$

The eigenvector corresponding to the largest eigenvalue is defined as the direction defining the *profile axis*, \mathbf{v} and is used in localized feature computations.

- The concavity metric is defined in order to distinguish between an active site that is relatively uniformly smooth and one that has many local undulations on its surface. The concavity is measured as the average distance between an active site surface atom and its closest n protein atoms. Since, the concavity metric is a measure of local undulations in the surface, n is chosen to be a relatively small number, in this case, 3. These local concavity values are then averaged to yield the surface concavity metric.

$$\Gamma(a_i) = \frac{1}{n} \sum_{j=1}^n \|a_i - b_j\| \quad (2.2)$$

$$\Gamma(S) = \frac{\sum_{i=1}^A \Gamma(a_i)}{A} \quad (2.3)$$

In these equations b_j is the j^{th} closest protein atom to active site surface atom

a_i and $A = |S|$.

- The curvature of a pocket defined as the spread of the pocket around its center of mass ($C_m(S)$) is calculated as:

$$\mathcal{K}(S) = \frac{\mu_p}{\sigma_p} \quad (2.4)$$

where

$$\mu_p = \frac{\sum_{i=1}^A \|a_i - C_m(S)\|}{A} \quad (2.5)$$

and

$$\sigma_p = \sqrt{\frac{\sum_{i=1}^A (\|a_i - C_m(S)\| - \mu)^2}{A - 1}} \quad (2.6)$$

are the mean and standard deviation, respectively, of the spread of the active site surface atoms around the center of mass of the site, $C_m(S)$.

- 3D invariant moments which are descriptors of geometric shape that are invariant to rotation and translation [112]. These invariants are calculated as follows:

$$\begin{aligned} J_1 &= \mu_{200} + \mu_{020} + \mu_{002} \\ J_2 &= \mu_{200}\mu_{020} + \mu_{200}\mu_{002} + \mu_{020}\mu_{002} - \mu_{110}^2 - \mu_{101}^2 - \mu_{011}^2 \\ J_3 &= \mu_{200}\mu_{020}\mu_{002} + 2\mu_{110}\mu_{101}\mu_{011} - \mu_{002}\mu_{110}^2 - \mu_{020}\mu_{101}^2 - \mu_{200}\mu_{011}^2 \end{aligned} \quad (2.7)$$

where

$$\mu_{pqr} = \sum_x \sum_y \sum_z (x - \bar{x})^p (y - \bar{y})^q (z - \bar{z})^r \quad (2.8)$$

and where \bar{x} , \bar{y} and \bar{z} are the coordinates of the center of mass of the active site surface.

In addition to these features that capture global variations in active site shape, novel cross-sectional features are also defined. These features are used for a finer-grained

characterization of the active site shape profile along the profile axis. The axis acts as a local frame of reference to place all the examples of the active sites in canonical positions and the features then capture the spatial variations in shape.

- Cross sections of the pocket at equal spacings (1\AA) from the center of mass, $C_m(S)$, and along the profile axis are considered. The distance between any two active site surface atoms in the cross-section is computed and averaged. The cross-section of active site surface S at distance r from the center of mass is defined as the set of vertices:

$$\Omega(S, r) = \{a_i \in S | \overline{a_i p_r} \perp \mathbf{v}\} \quad (2.9)$$

where \mathbf{v} defines the profile axis and p_r is a point on the profile axis that is r \AA away from the center of mass $C_m(S)$; $\|p_r - C_m(S)\| = r$.

Now the cross-sectional descriptor of the active site surface at a distance r can be described as the average pairwise distance among vertices in the cross-section:

$$\hat{\Omega}(S, r) = \frac{\sum_{i=1, j=1}^{M_r} \|c_i - c_j\|}{M_r(M_r - 1)/2} \quad (2.10)$$

for $c_i, c_j \in \Omega(S, r)$ where M_r is the total number of active site surface atoms in the cross-section $\Omega(S, r)$.

Additionally, electrostatic features are defined to capture the spread of charge and hydrophathy across the active site surface based on the electrostatic potential across the active site surface. The electrostatic potential at each active site surface atom is based on the partial charges of all the protein atoms. The partial charges used were the same as the ones used by AMBER [25] in their computation of the molecular mechanical force field to compute interaction energies.

The potential on an active site surface atom a_i due to a charge q_j placed at a distance d_j from it is given by:

$$V(a_i) = \sum_{j=1}^N \frac{q_j}{4\pi\epsilon_0 d_j} \quad (2.11)$$

where N is the total number of protein atoms and ϵ_0 is the absolute permittivity. While, in this study we use the Coulomb equation for potential calculations and do not consider the effects of solvent, this method can be extended using Poisson-Boltzmann solvers such as Delphi [55].

Based on the atom types used in [73], each protein atom was categorized as hydrophobic, hydrophilic or charged. This definition was then extended to define the hydropathy of the active site surface atoms based on the majority classification of its n closest protein atoms. Once again, in order to capture local information, a small number of closest neighbors ($n = 3$), is used.

The equation used to categorize the hydropathy (Y) of an active site surface atom a_i is as follows:

$$Y(a_i) = \text{majority}(Y(p_j)), j = 1 : n \quad (2.12)$$

where p_j is the j^{th} closest protein atom to a_i where $Y(p_j) \in \{H, P, C\}$.

The features used to capture the global chemical nature of the active site based on the previous definitions of charge and hydrogen bond propensity are:

- The global hydropathy features of the surface S measuring hydrophobicity, Y_H , hydrophilicity, Y_P and charge, Y_C are computed as follows:

$$Y_X(S) = \frac{\sum_{j=1}^A y_x(a_j)}{A} \quad (2.13)$$

where

$$y_x(a_j) = \begin{cases} 1 & \text{if } Y(a_j) = X \\ 0 & \text{otherwise} \end{cases} \quad (2.14)$$

The diverse patterns in active site chemistry are further captured using localized electrostatic features.

- Distribution of potentials across the active site surface: These features calculate the percentage of the active site surface occupied by positive, negative and neutral potential points respectively. The electrostatic nature of each active site surface atom is defined as follows:

$$E(a_i) = \begin{cases} P & \text{if } V(a_i) \geq \phi_1 \\ N & \text{if } V(a_i) \leq \phi_2 \\ O & \text{otherwise} \end{cases} \quad (2.15)$$

where ϕ_1 is set to 0.5 and ϕ_2 is set to -0.5 based on empirical observations of the variation of potentials in the example active site pockets.

The electrostatic spread features of the active site surface S measuring spread of positive potentials, Δ_P , spread of negative potentials, Δ_N and the spread of neutral potentials, Δ_O (by a given distance r) are calculated as follows:

$$\Delta_X(S; r) = \frac{\int_S \int_S u_r(a_i, a_j) \delta_X(a_i, a_j)}{(Area(S))^2} \quad (2.16)$$

where

$$u_r(a_i, a_j) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(r - \|a_i - a_j\|)^2} \quad (2.17)$$

and

$$\delta_X(a_i, a_j) = \begin{cases} 1 & \text{if } E(a_i) = E(a_j) = X \\ 0 & \text{otherwise} \end{cases} \quad (2.18)$$

where $X \in \{P, N, O\}$ and $\Delta(S; r)$ gives the average electrostatic property match

between all pairs of points on S (double integral) separated by a given distance, r (weighted by the Gaussian kernel u).

There are a total of 37 features: 16 geometric features and 21 electrostatic features:

$$\Phi(\mathbf{x}) = \langle \lambda_{1...3}, \Gamma, \mathcal{K}, J_{1...3}, \hat{\Omega}(r_1), Y_H, Y_P, Y_C, \Delta_P(r_2), \Delta_N(r_2), \Delta_O(r_2) \rangle \quad (2.19)$$

where $r_1 = 2...9\text{\AA}$ and $r_2 = 4...9\text{\AA}$.

Figures 9 and 10 show the variation of 2 of the above features for a subset of active site classes. These figures show that the feature-values between classes show significant overlaps. Figure 9 shows that pyridoxal active site pockets tend to be larger than those of the other 5 fragments. While, both pyridoxal and nicotinamide active site pockets seem to have very similar distributions of the largest eigenvalue, there are no nicotinamide pockets that are as large as some of the pyridoxal pockets.

In this study, the protonation states of the residues is not taken into consideration while analysing the active site electrostatics but this analysis can be included to better understand the active site chemistry. pK_a servers like $H++$ [44] can be used to approximate the protonation states of all of the protein residues before the electrostatic potential is computed at the active site. Additionally, the user can also specify the protonation states of the relevant residues as and when the information is available and this information can then be used in the electrostatic analysis.

D. Position-Dependent Features Based on Eigenvectors of the Moment of Inertia Matrix

Since the localized stereochemical features are rotation-invariant they cannot capture the positional variance in active site shape and chemistry. For *e.g.* there is no way to

Variation of the Largest Eigenvalue of the Coordinate Covariance Matrix

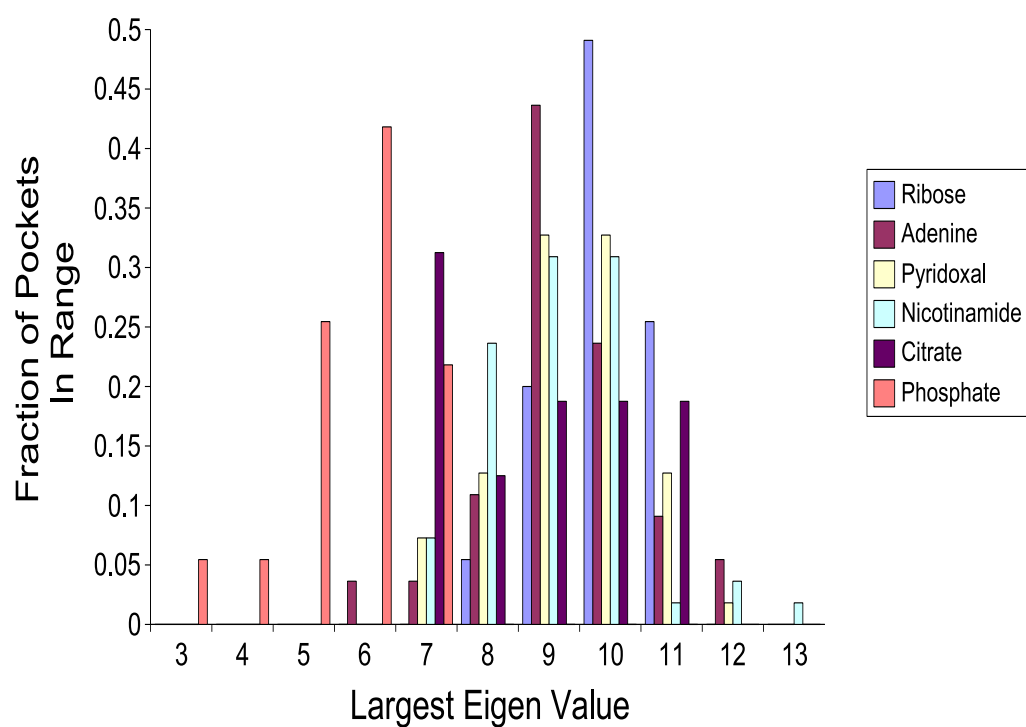


Fig. 9. The variation of the largest eigenvalue of the coordinate covariance matrix for the uniform-radius active sites belonging to six fragment classes

Cross-Sectional Shape Variation at 4Å

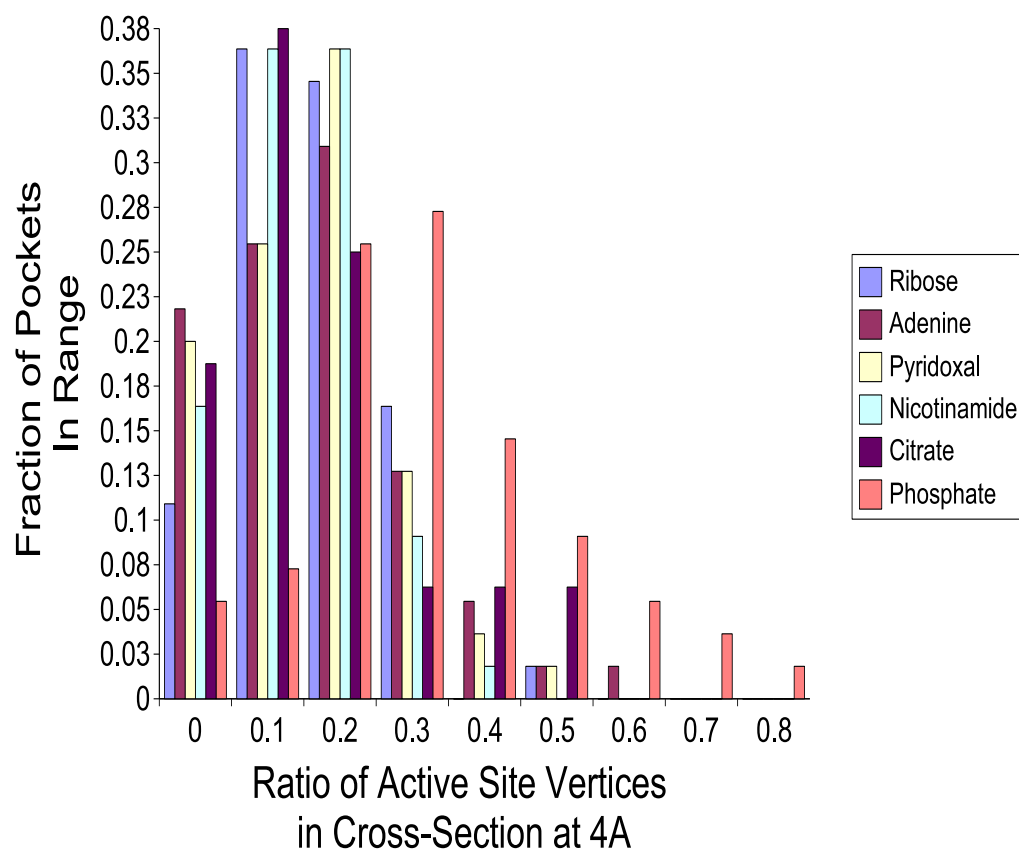


Fig. 10. The variation of the cross-sectional feature at 4Å for the uniform-radius active sites belonging to six fragment classes

capture the presence of a positively charged residue contacting the active site pocket right of the centroid at a distance of 3Å. The localized features will only be able to capture the distance of the positive charge but unable to pinpoint its direction from the centroid. This detailed and position-dependent description of the active site greatly enhances the possibility of accurately describing and recognizing the active site interaction patterns. To this end position-dependent features are developed.

For each active site pocket in the database, the eigenvectors of the moment of inertia matrix, \mathbf{I} , give the principal directions of the pocket in 3D space. These eigenvectors can therefore be used to compute feature vectors that capture the spatial distribution of geometric as well as the electrostatic patterns observed in the pocket w.r.t these principle directions. These feature vectors can then be used to compare different pockets and measure the similarities between various pocket shapes and electrostatics.

In order to compute these position-dependent features, the moment of inertia matrix, I , defined as

$$I = \begin{bmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{bmatrix}$$

is first computed where

$$\begin{aligned} I_{xx} &= \sum_{k=1}^N (y_k^2 + z_k^2) & I_{yy} &= \sum_{k=1}^N (x_k^2 + z_k^2) & I_{zz} &= \sum_{k=1}^N (x_k^2 + y_k^2) \\ I_{xy} &= - \sum_{k=1}^N x_k y_k & I_{xz} &= - \sum_{k=1}^N x_k z_k & I_{yz} &= - \sum_{k=1}^N y_k z_k \end{aligned} \quad (2.20)$$

and $\langle x_k, y_k, z_k \rangle$ represents the coordinates of the k^{th} active site atom. However, this definition of I assumes that the active site pocket is centered at the origin and therefore the actual atom coordinates $\langle X_k, Y_k, Z_k \rangle$ of each active site atom need to

be translated so that the centroid of the pocket is at the origin. This translation is defined as

$$A_T = \{\langle X_k - x_c, Y_k - y_c, Z_k - z_c \rangle\} \quad \forall k = 1 : K \quad (2.21)$$

where K is the number of active site pocket atoms, A_T represents the translation of the active site pocket to the origin, $\langle 0, 0, 0 \rangle$, obtained by subtracting the coordinate of the active site pocket centroid ($\langle x_c, y_c, z_c \rangle$) from each active site atom coordinate. This translated pocket is then rotated onto the reference framework as defined by the eigenvectors of I by multiplying the translated pocket coordinates by the eigenvector matrix as

$$A_R = A_T \cdot E \quad (2.22)$$

where E is the eigenvector matrix. Since the eigenvectors do not have directionality information, there are 4 possible representations of the eigenvectors given by

$$\begin{bmatrix} e_{xx} & e_{xy} & e_{xz} \\ e_{yx} & e_{yy} & e_{yz} \\ e_{zx} & e_{zy} & e_{zz} \end{bmatrix}, \begin{bmatrix} e_{xx} & e_{xy} & e_{xz} \\ -e_{yx} & -e_{yy} & -e_{yz} \\ -e_{zx} & -e_{zy} & -e_{zz} \end{bmatrix}, \begin{bmatrix} -e_{xx} & -e_{xy} & -e_{xz} \\ -e_{yx} & -e_{yy} & -e_{yz} \\ e_{zx} & e_{zy} & e_{zz} \end{bmatrix} \& \begin{bmatrix} -e_{xx} & -e_{xy} & -e_{xz} \\ e_{yx} & e_{yy} & e_{yz} \\ -e_{zx} & -e_{zy} & -e_{zz} \end{bmatrix}$$

These matrices correspond to 4 canonical rotations of the pocket along each of these eigenvector matrices. These 4 canonical pocket rotations are used to define the position-dependent features. For each canonical representation, the corresponding three eigenvectors are each divided into bins ranging from -3 to 3. These values were chosen since all the active site pockets are defined as uniform 5\AA radius patches. Additionally, since the pockets are translated to the origin, the maximal distance between the origin (pocket centroid) and any active site pocket atom has to be $\leq 5\text{\AA}$.

Therefore, the values of i , j and k are guaranteed to lie in the interval $[-3, 3)$. Each transformed active site atom is assigned a sector based on its coordinate value as

$$\langle i, j, k \rangle = \langle \lfloor \frac{x_k}{2} \rfloor, \lfloor \frac{y_k}{2} \rfloor, \lfloor \frac{z_k}{2} \rfloor \rangle \quad (2.23)$$

E. Geometric Similarity Analysis

The definition of sectors used in the previous section are used to capture the variation of spread of the pocket in coordinate space (measure of the geometric variation of shape). The final position-dependent feature vector for the active site pocket is computed as

$$F_s = \langle f_1 f_2 \dots f_K \rangle \quad (2.24)$$

where f_m refers to the m^{th} feature vector and is computed as

$$f_m = \text{count of active site atoms in the } m^{th} \text{ bin defined by indices } \langle i, j, k \rangle \quad (2.25)$$

In this study, the feature vector has $6^3 = 216$ features and each feature computes the spatial distribution of the pocket in a particular direction as defined by the moment of inertia vectors. When comparing the position-dependent feature vectors of two different active site pockets, it is necessary to compare the feature vectors for all the 4 canonical pocket rotations in order to make a fair comparison (since both the pockets may not be in the same orientation).

F. Dimensionality Reduction and Classification

All of the features described in Section C or those described in Section D need not have information equally relevant to classification. In this study, *Singular Value De-*

composition (*SVD*) is used to project the feature vectors onto the directions with maximum variability within the data. The *SVD* analysis can be used to reduce the dimensionality of the feature vectors and to increase the accuracy of classification of active sites. The *SVD* projections of the data are then analyzed using *linear Discriminant Analysis* (*LDA*). This technique is traditionally used to perform dimensionality reduction while retaining information pertinent to discrimination between classes by finding projections of feature data that yield the maximum class separability.

The combination of *SVD* and *LDA* projections has been used successfully to improve classification accuracy in many pattern recognition algorithms [13], [79]. The *SVD* dimensionality reduction selects for those feature-axes projections that best capture the variations in the data and thereby reduce the effects of noise and the *LDA* projections maximize the class-separability of the reduced-dimension features (output of *SVD* analysis). Both these techniques have complementary strengths and capture different information from the data and a combined technique enjoys an improved accuracy by combining the strengths of both the individual approaches.

1. Singular Value Decomposition

Given a $m \times n$ matrix \mathbf{M} that contains the feature vectors for the training data (such that m is the number of features and n is the number of training examples), the singular value decomposition of \mathbf{M} is given by:

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (2.26)$$

where \mathbf{U} and \mathbf{V} are unitary matrices that contain basis vectors describing the principal directions of variation in \mathbf{M} . The matrix $\mathbf{\Sigma}$ contains the singular values of \mathbf{M} which are weights for each of the directions of variation (right singular vectors in \mathbf{V} .) The

directions of variation are linear combinations of features from the feature space. The projection of training data onto SVD space helps to improve separation among clusters belonging to different fragment classes by emphasizing directions of high variation between classes. Additionally, in this space feature vectors that do not contain any information content have very low singular values (close to 0) and can therefore be ignored.

The transformed axes corresponding to the top 15 singular values were chosen for further computations when analyzing the localized stereochemical features and the transformed axes corresponding to the top 34 singular values were chosen when analyzing the position dependent features. In both cases, the number of dimensions used in further analyses was chosen such that enough dimensions were chosen so as to account for 90% of the variance in the data. Figure 11 shows how the top 15 singular values contribute to the majority of the information available in the localized stereochemical features.

In order to use this reduced dimension for classification, test example \mathbf{q} is also projected onto the lower dimensional SVD space as follows:

$$\mathbf{q}_{svd} = \mathbf{q}^T \mathbf{U} \mathbf{\Sigma}^{-1} \quad (2.27)$$

G. Linear Discriminant Analysis

The projections of feature data that yield the maximum class separability are given by the eigenvectors corresponding to the maximal eigenvalues of the matrix $\mathbf{S}_{\mathbf{W}}^{-1} \mathbf{S}_{\mathbf{B}}$ where $\mathbf{S}_{\mathbf{W}}$ is the within-class scatter matrix defined for a C-class problem as

$$\mathbf{S}_{\mathbf{W}} = \sum_{c=1}^C \sum_{\mathbf{x} \in \omega_c} (\mathbf{x} - \mu_c)(\mathbf{x} - \mu_c)^T \quad (2.28)$$

Singular Values Used to Discriminate Between Active Sites (Logarithmic Scale)

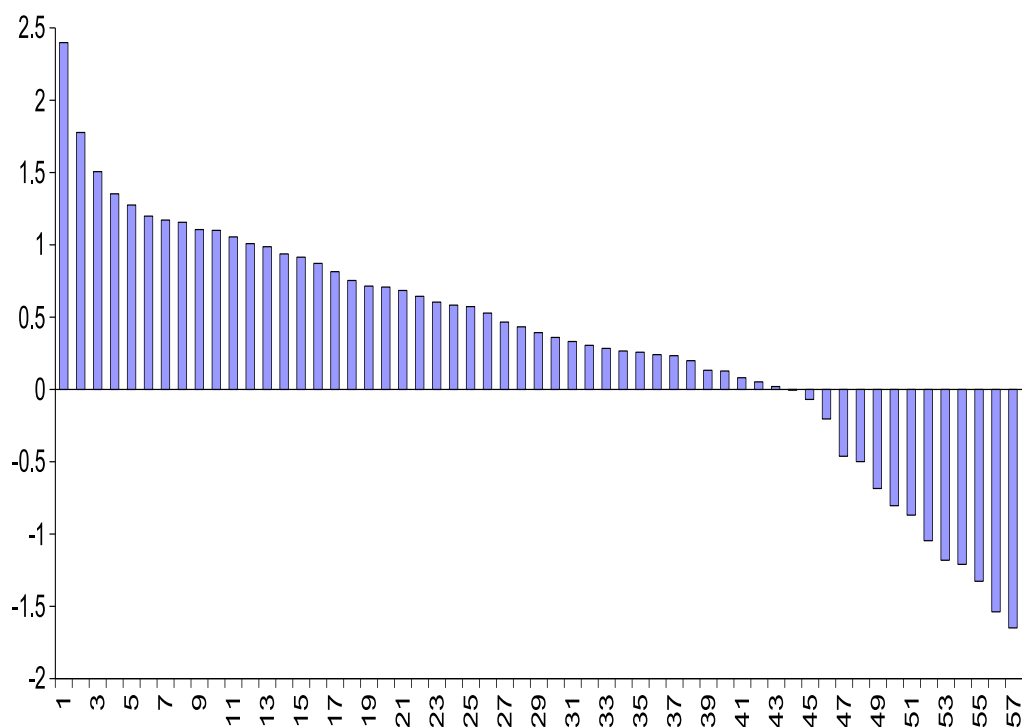


Fig. 11. The variation of the singular values obtained from an SVD analysis of the training active sites. It shows the significant variation in singular values and the relative importance of information in each of the transformed axes

where

$$\mu_{\mathbf{c}} = \frac{\sum_{\mathbf{x} \in \omega_c} \mathbf{x}}{N_c} \quad (2.29)$$

where \mathbf{x} is the feature vector, ω_c is the c^{th} class, N_c is the number of examples in class c and $\mu_{\mathbf{c}}$ is the mean vector over all the examples in class c (in this study, $\mathbf{S}_{\mathbf{W}}$ matrix is not weighed by the size of each class) and $\mathbf{S}_{\mathbf{B}}$ is the between-class scatter matrix given by

$$\mathbf{S}_{\mathbf{B}} = \sum_{c=1}^C N_c (\mu_{\mathbf{c}} - \mu)(\mu_{\mathbf{c}} - \mu)^T \quad (2.30)$$

where

$$\mu = \frac{\sum_{\mathbf{x} \in \omega_c} N_c \mu_{\mathbf{c}}}{N} \quad (2.31)$$

and N is the total number of examples and μ is the mean vector over all the examples in the dataset.

Given a set of examples from C classes, *LDA* yields $C - 1$ projections of the feature data given by the eigenvectors of $\mathbf{S}_{\mathbf{W}}^{-1} \mathbf{S}_{\mathbf{B}}$.

In the case of the localized stereochemical features, there are 441 fragment classes, but the feature data after SVD has only 15 dimensions and therefore all of the projections returned from *LDA* can be used in classification. The *LDA* projections in this case are not being used for dimensionality reduction but only to project the SVD-reduced feature space onto dimensions where class separation is maximum; allowing for maximum classification accuracy.

1. Classifier Based on Kernel Density Estimation

The classification of a test active site is based on the similarity between its reduced dimension feature vector to those in the database. In this study, the posterior probability of each fragment class given the observed test vector is computed using Kernel

Density Estimation. A *Product kernel* of D single-dimensional Gaussian kernels [32] is used to describe the spread of feature vectors in each fragment class in our database. This probability distribution is given by

$$P_{KDE}(\mathbf{x}|c_i) = \frac{\sum_{j=1}^N \frac{1}{h_1 \cdot h_2 \cdot h_3 \dots h_d} \prod_{d=1}^D K\left(\frac{x-x^j}{h_d}\right)}{N} \quad (2.32)$$

$$K\left(\frac{x-x^j}{h_d}\right) = \frac{1}{\sqrt{2\pi}h_d} e^{-\frac{1}{2}\left(\frac{x-x^j}{h_d}\right)^2} \quad (2.33)$$

where h_d gives the optimal bandwidth of each of the Gaussian kernels and is determined using $h_{opt} = 0.9AN^{-\frac{1}{5}}$ where N is the number of examples in the class being considered. A is defined as $A = \min(\sigma, \frac{IQR}{1.34})$ where IQR is the *Interquartile Range* for that particular dimension and σ is the sample deviation.

Assuming equal prior probabilities for all fragment classes, this probability density function can be used to estimate the likelihood of a class C_i given a test feature vector ($\Phi(\mathbf{x})$) as $(P(C_i|\Phi(\mathbf{x})))$. This probability is used as an indicator of confidence in the class prediction based on the feature analysis.

The single-dimension Gaussian kernels used in the product kernel assume that the individual features are all independent of each other. It should be noted here that while the projections returned by *SVD* are independent dimensions, those returned by *LDA* analysis are not necessarily independent. In this study, we make the assumption that the use of independent kernels is sufficient to accurately represent the feature space of active site descriptions.

H. Conclusions

In this chapter, two different systems of features, both of which are designed to capture the diversity in interaction patterns observed within active sites that bind the same ligand have been presented. These feature systems go beyond the traditional

global descriptions of active site stereochemistry and capture local variations in the stereochemistry of the active site thereby enabling greater accuracy in the description and comparison of active site pockets.

The traditionally well-accepted dimensionality reduction techniques of *SVD* and *LDA* are used to find features that have the most variability in information as well as those that contain information relevant to discrimination between various fragment classes. Chapter V analyzes the results of these methodologies and examines the classification power of these features.

CHAPTER III

COMBINATION OF INDIVIDUAL FRAGMENTS

The main motivation for the fragment-based analysis presented in this dissertation was to address the flexibility observed in ligands with multiple rotatable bonds. Since the classification scheme proposed in Chapter II identifies individual fragments, the final analysis of the enzymatic function of a protein requires that these individual fragment classifications be combined to yield one multi-fragment ligand class (for example, combining fragment classification of *adenine*, *ribose* and *phosphate* to yield an overall final classification of AMP). In this chapter, an algorithm based on *Markov Random Field* theory [74] to combine these individual fragment classifications into one single multi-fragment ligand classification will be presented. The overall procedure for analyzing an apo protein based on this algorithm is presented in Section A. Section B details the theory behind the *Markov field* designed to combine individual classifications. Section C characterizes the flexibility observed in each of the multi-fragment ligands in the database.

A. Analyzing Active Site Pockets for the Multi-Fragment Ligands

For each of the protein complexes in our database bound to any of the multi-fragment ligands in this study, a 10 Å uniform pocket centered at the centroid of the ligand was defined. This pocket defines the active site for the multi-fragment ligand throughout the rest of this study. For each of the mesh points in these active site pockets, a uniform 5Å patch centered at that point is generated. A large active site with N mesh points is therefore represented by N uniform 5Å patches. These patches overlap and multiple patches describe the various regions in the larger active site. These patches are input to the fragment classification algorithm in order to identify

various fragments that might fit into the larger active site.

The possible labels for each mesh point is the set of all fragment class labels. Each class label is associated with a probability value based on the posterior probability obtained from the feature-based classification for the uniform patch centered at that mesh point. It is possible that some class labels have higher probabilities while others have zero probabilities but they all sum to 1.

$$Label(m) \in F \cup \perp \quad (3.1)$$

where $Label(m)$ is the label for the m^{th} mesh point and F is the set of all fragment class labels and \perp represents the NULL class s.t.

$$\sum_{i=1}^C P_m(f_i) = 1 \quad (3.2)$$

where $P_m(f_i)$ is the posterior probability obtained from classification of the m^{th} mesh point for the i^{th} fragment class label and C is the number of fragment classes in the database plus the NULL class.

All the fragment classes associated with non-zero posterior probabilities represent possible fragments centered at that mesh point. Therefore, each mesh point in the large active site pocket has at most C fragment labels that represent the fragments that could be bound centered at that mesh point. For a large active site with N mesh points, at most C^N possible label combinations exist based on equation 3.1. For a ligand like *AMP*, the correct labeling of the mesh points in the active site would be *NULL* for all but 3 mesh points, closest to the centers of each fragment, which are labeled as *adenine*, *ribose* and *phosphate*. Finding this combination of labels from the list of all possible labels so as to identify the large ligand that binds the site is extremely difficult since enumerating each of the possible label combinations is

time-consuming and unrealistic. Therefore a recombination scheme based on *Markov random field* was developed. The goal of the recombination scheme is to find the joint probability of binding a ligand given the probability of binding individual fragments as well as the additional contextual constraints (provided by the geometric relationship between ligand fragments) to find a labeling for all mesh points in the active site that is most consistent with binding at individual sites.

B. Markov Random Field Theory

Markov Random Field (*MRF*) theory has been used in image-processing algorithms as a way to model context-dependent information [42], [133]. *MRFs* seek to characterize mutual influences amongst variables by using conditional probability distributions. It is computationally expensive to model and compute the contextual dependencies between all the variables in a system. This is especially true in this application as computing the joint probability of various active site mesh points being assigned a set of labels and analyzing all possible combinations is computationally intractable. *MRFs* provide one avenue for analyzing the relationships between all possible labels by limiting the number of dependencies to be considered based on the *Markov* assumption (only neighbors of a site have an effect on its labels) and the development of various sampling techniques have enabled to make this analysis more tractable.

1. Formulating as a Labeling Problem

MRFs have often been used to study labeling problems, and in this study the combination of fragments problem will be formulated as a labeling problem. In this labeling problem, a set of sites, S (active site mesh points), can take on a set of labels, C (fragment class labels). The goal of labeling problems is to find a labeling of sites that is

most consistent with the observations at these sites. In this study, this translates to finding a labeling of sites with fragment class labels that is most consistent with the stereochemical feature patterns observed at each of these mesh points.

Configuration Space: In *MRF* theory the set of all possible labeling of sites is referred to as configuration space. In our problem, since each of the mesh points can take on any of the C labels, the configuration space is the set of all possible labels ($C^{|S|}$). A typical active site can contain anywhere between 300-500 mesh points and in this application, there are 441 fragment classes plus the NULL class and therefore the size of our configuration space is approximately 442^{400} .

Discrete Labels: In any labeling problem, the labels could either be continuous or discrete. In this study, we have discrete labels and the labeling of any site in S is given by $f = \{f_1, \dots, f_k\}$ where f_i s are the fragment class labels. The labels are categorized this way to enable comparisons of similarity between labels during analysis. In this study, since all ligand fragments with *Tanimoto* score > 0.7 have already been clustered, this issue of similarity between classes will not play a major role. But this notion of similarity could have been used in place of the clustering approach used in this study.

2. Local Neighborhoods to Evaluate Contextual Information

The full joint probability of the labeling of all the mesh points in the active site is written as

$$P(f) = P(f_1|f_2, f_3, \dots, f_n)P(f_2|f_1, f_3, \dots, f_n) \cdots P(f_n|f_1, f_2, \dots, f_{n-1}) \quad (3.3)$$

The above equation assumes that the labeling at each site depends on all other sites in the active site. In practice, defining these conditional probabilities is impossible and unnecessary. The other option would be to assume that labeling of the mesh points

were independent of each other. Then the probability of observing any particular labeling of the sites is given can be written as

$$P(f) = \prod_{i \in S} P(f_i) \quad (3.4)$$

where $P(f_i)$ are the individual class probabilities at each site, *i.e.* a *probability distribution function* over the various class labels (estimated posterior probabilities $p_i(f|\Gamma)$ where Γ is the feature vector for the patch centered at mesh point i). The above equation seeks to find the optimal labeling for the entire active site by finding the best label at each site. Unfortunately, this strategy fails to account for interactions between some of the labels within the site and these contextual dependencies are not capture in Equation 3.4. Therefore, neither Equation 3.3 nor Equation 3.4 completely capture the interactions between the mesh point labels.

In this application, each multi-fragment ligand is formed by the placement of various fragment classes in a specific configuration. There are constraints on the placement of these fragments which lead to correlations between the stereochemical features observed at neighboring sites and hence their labeling. For example, in order to label an active site as binding the ligand *AMP*, it is not only necessary to find 3 mesh points labeled *adenine*, *ribose* and *phosphate*, but, it is also necessary that these mesh points be placed in 3D space such that they satisfy the geometric constraints based on bond distances and bond angles observed in typical *AMP* conformations. The simplified relationship between the various mesh points described in equation 3.4 is unable to capture these spatial constraints and it is necessary to calculate the conditional probabilities between the observations of the labels at various mesh points in order to fully describe the system. Due to the large size of the configuration space, enumerating all of these conditional probabilities is not feasible. Therefore, a *MRF* is used to incorporate the contextual dependencies as conditional probabilities.

Neighborhood: The neighborhood of any site is the set of all other sites that have an influence over its labeling. It is often defined as all sites within a radius r , *i.e.*,

$$N_i = \{i' \in S \mid \text{dist}(i, i') \leq r, i \neq i'\} \quad (3.5)$$

This definition of neighborhoods reduces the number of contextual dependencies that need to be modeled in order to clearly describe/analyze the system since the assumption is that the neighborhood of a site has greater effect on its label than other sites that are not within the neighborhood. This definition of neighborhoods is also referred to as *Markov mesh models* or *Markov Blankets*. This conditional independence is formulated as

$$P(f_i | f_{S-\{i\}}) = P(f_i | f_{N_i}) \quad (3.6)$$

In this study, two neighborhood definitions are used. The first definition of neighborhood for a mesh point i , N_i , is geared towards the combination of fragments into ligands and therefore is based on the geometric constraints between fragments of a ligand. Typically, the distance between fragments of a ligand is at most 7Å and therefore a r value of 7Å was used to define neighborhoods. For *e.g.*, when analyzing an *AMP* site, the neighborhood for a mesh point labeled *adenine* is a set of all other mesh points that are within 7Å of it. This is because in all observed conformations of *AMP*, the adenine fragment is placed next to the *ribose* fragment and the maximum distance between the centroids of these two fragments is approximately 7Å. The assumption of non-interaction between non-neighboring mesh points is met since 3 mesh points labeled *adenine*, *ribose* and *phosphate* can constitute the ligand *AMP* as long as the spatial relationship between these points is satisfied and neither the labeling of other sites nor the geometric placement of other sites affects the probability of the fragments of ligand *AMP* being placed at these three mesh points. Therefore,

it is possible to model the neighborhood distance constraints based on their mutual geometric constraints.

The second definition of neighborhood for a mesh point i , N'_i , is based on the nature of each ligand fragment. If a fragment is placed centered on mesh point i , it occupies the active site such that it covers neighboring mesh points and no mesh point within a radius r (defined by the radius of the fragment) can be labeled as the center of any other ligand fragment since otherwise fragments would overlap. This constraint on fragment center positions adds an additional layer of neighborhood constraints on the labellings such that if mesh point i is labeled as f_i it reduces our belief that another mesh point i' immediately adjacent to i would be labeled anything other than f_i . Therefore the second neighborhood is defined by a radius of 2\AA the minimum distance between any two fragment centers.

Both these neighborhood definitions are symmetric *i.e.*, if mesh point i' is in the neighborhood of mesh point i , it follows that mesh point i is in the neighborhood of mesh point i' since the above definitions of neighborhoods are based on the *Euclidean* distance metric. Using the neighborhood definition, the probability of observing any particular labeling of the sites can be written as

$$P(f) = \prod_{i \in S} P(f_i | f_{N_i}) \quad (3.7)$$

Cliques: To further reduce the complexity of equation 3.7, conditional dependencies on all neighbors can be approximated by decomposing the neighborhood into cliques of various sizes. A clique, θ , for a given (S, N) is a subset of sites in S such that each is a neighbor of the other. A clique could contain just a single site ($\theta = \{i\}$) or could contain a pair of neighboring sites ($\theta = \{i, i'\}$). The set of all single site cliques is denoted as Θ_1 , and the set of all pairwise site cliques is denoted as Θ_2 etc. Therefore $\Theta = \Theta_1 \cup \Theta_2 \cup \Theta_3 \cdots$ represents the set of all possible cliques. The

cliques allow for the specification of the interactions that will be modeled within a system. Choosing to examine all cliques of size two or less implies that only single site interactions and pairwise interactions are considered to be important to completely defining the contextual dependencies within a give system.

Applying contextual constraints on labels two at a time is the lowest-cost constraint (computationally) and in practice, this also captures most of the dependencies between various labels in most applications [73]. This directly translates to considering all cliques of size 2 or less within a system. Therefore, in this study, only cliques of size 2 or less will be considered (considering the interaction between a fragment as well as the interactions between ligand fragments taken two at a time).

The formal definition for Θ_1 and Θ_2 is as follows:

$$\begin{aligned}\Theta_1 &= \{i|i \in S\} \\ \Theta_2 &= \{(i, i')|i \in N_{i'}, i' \in N_i\}\end{aligned}\tag{3.8}$$

The definition of cliques helps further simplify the joint probability defined in Equation 3.7 by considering the interactions within a neighborhood one or two at a time and the joint probability can now be written as

$$P(f) = \prod_{i \in S} P(f_i) \prod_{j \in N_i} P(f_i|f_j)\tag{3.9}$$

where N is the number of mesh points. Despite this simplification, maximizing this joint probability is still intractable in practice since these conditional probabilities are not often explicitly known but they can only be scored as quality/energy of interaction. For example, the contextual dependence between two labels can be determined by examining how well the distance between the corresponding mesh points fits the target profile of actual distances between those fragments. Therefore *MRF* problems are formulated as *Gibbs Random fields* since this provides a simple way to specify the

joint probability of any given labeling in terms of energy of different interactions. The equivalence between an *MRF* and a *GRF* was proved by Besag *et.al* [12]. A Gibbs distribution is written as

$$P(f) = Z^{-1} e^{-\frac{U(f)}{T}} \quad (3.10)$$

where

$$Z = \sum_{f \in F} e^{-\frac{U(f)}{T}} \quad (3.11)$$

is called the partition function and serves as a normalizing constant, $U(f)$ is the energy function which is defined as the sum of clique potentials over all possible cliques, Θ and T is the temperature (usually assumed to be 1; can be varied in order to control the sharpness of the energy function).

$$U(f) = \sum_{\theta \in \Theta} V_{\theta}(f) \quad (3.12)$$

where V_{θ} is the clique potential for the θ^{th} clique. According to this definition, the lower the energies, higher the probability of the occurrence of labeling f . As mentioned earlier, a simple way to convey contextual constraints as well as reduce complexity and analysis time is it do so with two labels at a time. This is done by considering only cliques of size two or less. Therefore, $U(f)$ can be written as

$$U(f) = \sum_{i \in S} V_1(f_i) + \sum_{i \in S} \sum_{i' \in N_i} V_2(f_i, f_{i'}) \quad (3.13)$$

where V_1 is the clique potential for all cliques of size 1 and V_2 is the clique potential for all cliques of size 2. In this study the single site interactions are captured by the classification algorithm based on the stereochemical features at each site. The strength of these interactions is available as the posterior probability of any given label at each mesh point and therefore this interaction need not be modeled as an

energy function. Therefore, Equation 3.10 can be rewritten as

$$P(f) = \prod_{i \in S} P(f_i) Z^{-1} e^{-\frac{\sum_{i \in S} \sum_{i' \in N_i} V_2(f_i, f_{i'})}{T}} \quad (3.14)$$

Therefore, the total joint probability of a labeling f given the feature matrix Γ_S (rows of the matrix contain Γ_i s, feature vectors at sites in S) is given by

$$P(f|\Gamma_S) = \prod_{i \in S} P(f_i|\Gamma_S) Z^{-1} e^{-\frac{\sum_{i \in S} \sum_{i' \in N_i} V_2(f_i, f_{i'}|\Gamma_S)}{T}} \quad (3.15)$$

For our application, the goal is to find a ligand L which has the maximal probability over all possible sites in the active site given the stereochemical features computed for the site, *i.e.* find

$$\Psi(S) = \arg \max_L \arg \max_f P(f|L, \Gamma_S) \quad (3.16)$$

where $\Psi(S)$ is the final ligand class for site S and $P(f|L, \Gamma_S)$ can be defined by rewriting Equation 3.15 as:

$$P(f|L, \Gamma_S) = \prod_{i \in S} P(f_i|L, \Gamma_S) Z^{-1} e^{-\frac{\sum_{i \in S} \sum_{i' \in N_i} V_2(f_i, f_{i'}|L, \Gamma_S)}{T}} \quad (3.17)$$

$P(f_i|L, \Gamma_S)$ is the posterior probability of a single site i being labeled as f_i given the stereochemical features and ligand L and $V_2(f_i, f_{i'}|L, \Gamma_S)$ is the pairwise interaction potential between sites i and i' labeled as f_i and $f_{i'}$ respectively given the stereochemical features and ligand L . There are two pairwise interaction potentials based on the two neighborhood definitions described above. In this application, the posterior probability depends only on the stereochemical features computed at each site and the type of ligand has no effect on this probability. Therefore L can be factored out and the single site clique potential can be rewritten as:

$$P(f_i|L, \Gamma_S) = P(f_i|\Gamma_S) \quad (3.18)$$

Similarly, in the case of the pairwise-interaction potentials, these values are not affected by the stereochemical features at the sites but are only based on the interaction between those sites as defined by geometric constraints between those sites. Therefore Γ_S can be factored out and the pairwise clique potential using both neighborhood definitions can be rewritten as:

$$V_2(f_i, f_{i'}|L, \Gamma_S) = V_2(f_i, f_{i'}|L) \quad (3.19)$$

The potential V_2 captures the interaction between pair-wise sites in all the cliques. In this study, this potential is defined as:

$$V_2(f_i, f_{i'}|L) = \Lambda(f_i, f_{i'}|L)\Upsilon(f_i, f_{i'}|L) \quad (3.20)$$

where Λ is the pair-wise potential due to geometric constraints between ligand fragments, *i.e.* based on the neighborhood definition N_i and Υ is the pair-wise potential due to geometric constraints based on the occupancy of a single fragment, *i.e.* based on the neighborhood N'_i .

Using the neighborhood definition N_i this pairwise relationship is modeled by analyzing the geometric distance relationships between ligand fragments taken two at a time. Each ligand occurs in nature in various conformations (due to ligand flexibility) with multi-fragment ligands having greater number of conformations. One way to capture the geometric constraints between fragments in a multi-fragment ligand is by capturing the distance between them. In each conformation, there is variation in the distance between the centers of the fragments in the ligand. Modeling these variations will allow for the definition of constraints on the contextual relationship between the fragment centers and therefore the labellings of the mesh points in the active site. In this study, the mean (μ) and the standard deviation (σ) of the distances are computed and the variation in distance between fragment centers of a given ligand

is modeled using a harmonic approximation as the square of the normalized deviation from the mean (assuming energy or cost of deviation scales up quadratically). Therefore the potential Λ given a ligand L , for 2 sites i and i' at a distance of $d_{ii'}$ from each other, can be computed as follows:

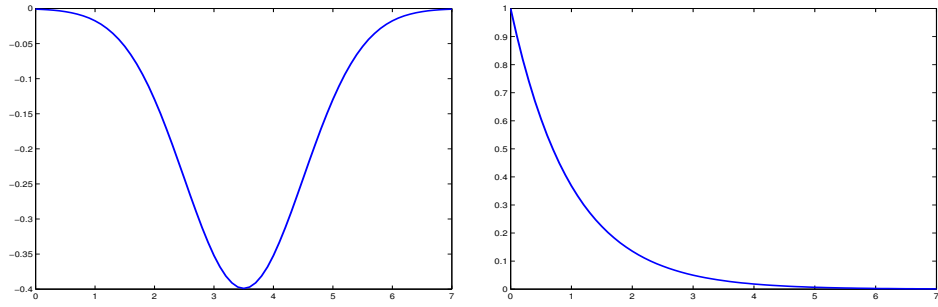
$$\Lambda(f_i, f_{i'}|L) = \begin{cases} -e^{\left(\frac{d_{ii'} - \mu(L, f_i, f_{i'})}{\sigma(L, f_i, f_{i'})}\right)^2} & \text{if } 2 < d_{ii'} < 7 \\ 1 & \text{if } d_{ii'} \leq 2 \end{cases} \quad (3.21)$$

Using the neighborhood definition N'_i , this pairwise interaction potential is measured by analyzing the spreads of each of the ligand fragments *i.e.*, for a given ligand fragment centered at mesh point i . Since each fragment occupies space within the active site, it is not possible for 2 fragment centers to be immediately adjacent to each other. Therefore there has to be a penalty if two sites immediately adjacent to each other, i, i' , are labeled differently and this penalty is a function of the confidence in the different label at i' , *i.e.* if the confidence in the different labeling at i' is low it has little or no effect on the confidence of the labeling at i whereas if the different labeling has high confidence, that decreases the confidence in the labeling at i . This penalty can be modeled as a function of the variation in distance between sites i, i' , $d_{ii'}$ from the actual minimum distance between the fragments at i and i' as follows:

$$\Upsilon(f_i, f_{i'}|L) = \begin{cases} e^{(r-d_{ii'})P(f_{i'})} & \text{if } d_{ii'} \leq 2 \\ 1 & \text{if } 2 < d_{ii'} < 7 \end{cases} \quad (3.22)$$

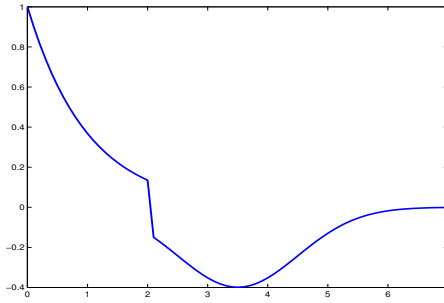
where r is the minimum distance between fragments f_i and $f_{i'}$ in ligand L . Figure 12 shows the graphical representation of the functions Λ , Υ and V_2 respectively.

Additionally, Figure 13 better illustrates the concept of these neighborhoods. Let us assume that the active site pocket depicted by the mesh in the figure is being analyzed for the probability that it binds the ligand *pyridoxal phosphate* (PLP) with



(a) Graphical Representation
of Λ

(b) Graphical Representation
of Υ



(c) Graphical Representation
of V_2

Fig. 12. Graphical representation of interaction potential functions defined in this study

two fragments *pyridoxal* and *phosphate*. If the center of the fragment *pyridoxal* is placed at mesh point labeled i , then it is impossible for the second fragment of *PLP* to be placed at mesh point j since there would be steric conflict between the two fragments. The interaction potential Υ takes into account these possible overlaps between ligand fragments and penalizes fragment combinations that show these steric conflicts. But, there would be no steric conflict if the center of the fragment *phosphate* would be placed at mesh point k and additionally the distance between mesh points i and k is within the range of observed distance between the fragments *pyridoxal* and *phosphate* in all conformations of *PLP*. This interaction between i and k is taken into account by the interaction potential Λ . Similarly, mesh point l is placed further away from mesh point i and the distance between these two fragments is not within the range of observed distances between the two fragments, thereby decreasing the probability that the ligand *PLP* is bound to the mesh such that the fragment *pyridoxal* is centered at i and the fragment *phosphate* is centered at mesh point l . The upper distance cutoff of 7Å for Λ takes this interaction into account. This upper distance cutoff was obtained based on empirical observation of the distances between fragments of ligands in this study (experiment detailed in Section C).

The fragment class labeling of the mesh points is based on the stereochemical features calculated at each mesh point. Nearby mesh points are more likely to have similar interaction patterns and therefore similar feature vectors. Therefore, it is more likely that nearby mesh points will have similar labellings. In an analysis of active sites binding multi-fragment ligands, we found that while the feature difference between the feature vectors of the actual fragment pocket and those of the subpocket centered at the closest mesh point was the least, nearby mesh points (those within 1.5-2Å from the fragment centroid), also had very small feature differences. Figure 14 shows this distribution and the lower distance cutoff of 2Å for Υ was based on this experiment.

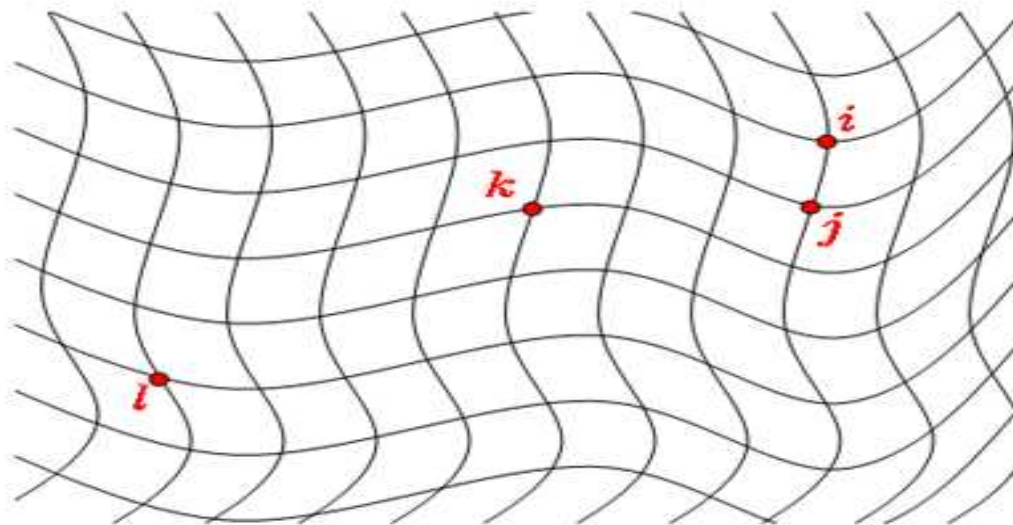


Fig. 13. Graphical representation of active site mesh to depict intuitively that it is impossible to place the centers of the two fragments of ligand *PLP* at mesh points *i* and *j* without causing steric conflict between these fragments. At the same time, it is quite possible that the fragments are placed at mesh points *i* and *k*. Additionally, given the geometric constraints that exist between the placement of *PLP* fragments, it is impossible for one of the fragments to be placed at mesh point *i* and the other to be at mesh point *l*.

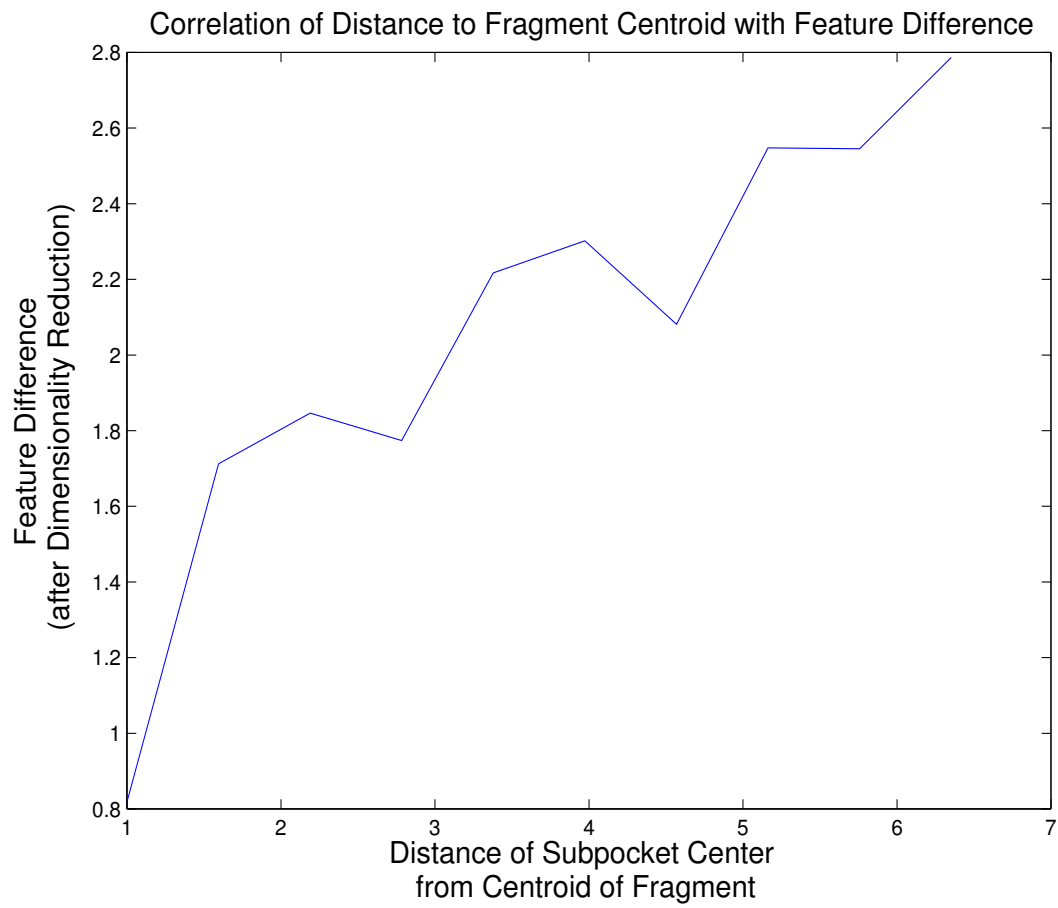


Fig. 14. The variation of feature difference with distance from actual fragment center for all the fragment classes in this study

C. Parameter Estimation

In practice, the values of μ and σ between various fragments of various ligands are computed by first generating various conformers for each of the multi-fragment ligands were generated using *Omega* [17] a module of the OpenEye software. *Omega* evaluates the geometrical restraints (bond angles and distances) experienced by any given chemical compound and generates possible conformations of the compounds based on these geometrical restraints. Each of the generated conformations is evaluated for steric as well as energetic favorability and conformations without steric conflicts and those conformations with energies within 50 kCals/mol of the lowest energy conformation and 1 Å r.m.s.d from previously sampled conformations are chosen for future analyzes. The number of conformations generated for a subset of the multi-fragment ligands is listed in Column 3 of Table I. This table shows the large variation in flexibility of each of these ligands (Pyridoxamine-5'-Phosphate (PMP) has only 6 possible conformers as opposed to Nicotinamide-Adenine-Dinucleotide Phosphate (NAP) which has 7000 conformers). This table also lists the number of rotatable bonds (obtained from *Pubchem* [135] in Column 2) for each of these ligands. The table shows that while there is a linear correlation between the number of rotatable bonds and the number of possible conformers in most cases, there are a few outliers. In cases where there are a large number of rotatable bonds but not equivalently large number of conformers, it is due to the fact that some of the single bonds are within a large ring system thereby reducing its flexibility. Similarly, in some cases, a large number of conformers is observed for a compound with fewer rotatable bonds due to the isomerism observed in the structure.

Table I.: The number of rotatable bonds and number of conformers generated by Omega for a subset of the multi-fragment ligands

Ligand	Number of Rotatable bonds	Number of Conformers Generated by Omega
2DT	4	39
AMP	4	52
ADP	6	554
ANP	8	1418
ATP	8	803
FMN	7	51
PMP	4	6
PLP	3	4
SAH	7	580
SAM	7	222
TDP	8	1417
TMP	4	36
UD1	11	1200
UDP	6	426
UPG	9	1590

Once the conformations are generated, each conformer was split into its constituent fragments and the center of mass was computed for each fragment and the distances between the centroids of each pair of fragments was computed. The means and the standard deviations of these distances over all the ligand conformers were then computed as:

$$\mu(L, f_i, f_{i'}) = \frac{1}{K} \sum_{i=1}^K d_{c_i, c_{i'}} \quad (3.23)$$

$$\sigma(L, f_i, f_{i'}) = \sqrt{\frac{\sum_{i=1}^K (d_{c_i, c_{i'}} - \mu(L, f_i, f_{i'}))^2}{K - 1}} \quad (3.24)$$

where K is the number of conformers for ligand L , $\mu(L, f_i, f_{i'})$ is the mean of the distances $d_{c_i, c_{i'}}$ between the centers of fragments f_i and $f_{i'}$ of ligand L over all the

conformers of ligand L and $\sigma(L, f_i, f_{i'})$ is the standard deviation of these distances.

D. Simulated Annealing Algorithm to Sample Conformational Space of Labels

Equation 3.17 provides a way to quantify the 'goodness' of a given labeling but the problem of searching the label conformational space for a labeling with the maximum probability still remains. In this section, a methodology to sample this space based on *Simulated Annealing* (SA) techniques is presented. Simulated annealing is a technique used in optimization problems to find a final solution that minimizes a given energy function [64]. While this technique does not guarantee a globally optimal solution, it is guaranteed to find a good solution even in the presence of noisy data. The SA procedure has its basis in the process of annealing solids practiced in metallurgy where the materials are initially heated to a temperature so as to allow for atomic rearrangements. When the atoms are heated, they are disordered leave their current energy states and randomly explore higher states of energy and the cooling allows the atoms to settle down in a lower energy state. Simulated annealing avoids getting stuck in local minima by sometimes allowing movement from a lower energy state to a higher energy state. This results in the sampling of larger regions of the state space and gives the algorithm the power of backtracking. SA techniques have often been used to solve *NP-complete* optimization problems like the *traveling salesman* problem [21]. Additionally, authors in [42] also used this sampling technique to find the best labeling in their application of *Markov random fields* to the identification of battlefield entities. In this study, SA will be used to find the labeling with the lowest energy. The pseudocode for the SA procedure used here is as follows:

Algorithm 1: Pseudocode for sampling the label conformational space using

Simulated Annealing

Input: Fragment class labels at each mesh point in active site and the euclidean distances between mesh points

Output: A labeling with the largest value of $P(f|L, \Gamma_S)$ over all sampled labellings

SAMPLING USING SA(1)

- (1) num cycle \leftarrow 0 Current labeling $C \leftarrow f'$
- (2) Compute $E' = P(f'|L, \Gamma_S)$ for f'
- (3) Generate new labeling f'' by randomly changing one or more of the labels in f' and/or by changing one or more of the mesh points being considered in f'
- (4) Compute $E'' = P(f''|L, \Gamma_S)$ for f''
- (5) **if** $E'' > E'$ **then** $C \leftarrow f''$
- (6) **else** Accept new labeling with probability $P(\frac{E''-E'}{kT})$
- (7) num cycle = num cycle + 1
- (8) **if** num cycle $> N$ **then** stop
- (9) **else** continue

where N is the total number of cycles for simulated annealing and in this study was chosen to be 100000. While most movements in the label conformational space are made to follow the energy gradient *i.e.* only accept a move if energy at new position is lower than the energy at current position, this algorithm allows for some moves where the energy is not as good as the current energy (based on the probability of observing the energy difference between the 2 states). This allows for the technique to escape local minima and allow for some backtracking. Additionally, when the algorithm observes a local minima, the next labelings considered are obtained by

exploring mesh points in the immediate neighborhood of the current mesh points. This allows us to explore small optimizations to the position of the ligand fragments in the active site that might better account for the distance relationships between the fragments forming the multi-fragment ligand.

E. Combining Probabilities from Multiple Models into a Unified Prediction of Most Likely Ligand

The final aim of this study is to combine the predictions across the entire active site to pick the best ligand, but the probabilities are not comparable in a fair way. This is due to the different number of constraints for different ligands based on the number of fragments that they contain. For example, a ligand with n fragments can be identified if and only if n active site mesh points are labeled correctly with the ligand fragments and the distances between these n mesh points also meet the geometric constraints between the fragments of the ligand. This is not the case of a single fragment ligand where the only requirement is the accurate classification of the fragment based on the stereochemical features. This makes it more likely that even when an active site binds a multi-fragment ligand, there is a greater probability that single-fragment ligands have a higher *MRF* joint probability, making a ranking based on the *MRF* probabilities unfair to multi-fragment ligands.

One possible solution is to base the ligand identification on the active site size. In the case of small ligands that contain only a single fragment, the steric analysis used by the *MRF* formulation is not applicable since inter-fragment spatial constraints do not exist. In a recent study of active site shapes by Kahraman *et. al.* [60] found that active sites that bind smaller ligands can be clearly differentiated from those that bind larger ligands by just comparing active site sizes. Based on this

study, assuming that the protein active sites that bind single-fragment ligands are smaller in size in order to increase protein-ligand interaction specificity, this size difference in active sites can be taken into account in the analysis. Active sites can be categorized based on size and the larger active sites can be analyzed using fragment identification using feature-based analyses and the fragments can be combined using the *MRF* formulation whereas the smaller active sites can be analyzed using only the feature-based analysis to find the single-fragment ligand that best fits the active site stereochemistry.

The other, more general solution is to normalize the joint probabilities from *MRF*. Ligands with larger number of fragments tend to have lower joint probabilities as opposed to those with only 2 fragments. Purely by chance, it is easier to find a single mesh point with a particular classification label than it is to find multiple mesh points with the correct classification labels as well as satisfying the specified geometric constraints. To avoid this unfair advantage to ligands with fewer fragments, it is necessary to normalize out these differences.

The final probability of identifying a single-fragment ligand is equal to the probability obtained from the feature-based classification methodology since there are no additional distance criteria to be applied in the *MRF* analysis. There are additional distance constraints applied to the analysis for multi-fragment ligands. This can be done by normalizing the *MRF* probability by the probability of observing any n -points (where n is the number of fragments in the ligand) that match the specified distance constraints. The normalization for a ligand with two fragments can be written as:

$$P'(f|L, \Gamma_S) = \frac{P(f|L, \Gamma_S)}{\sum_{i \in A_S} \sum_{j \in A_S} P(a_i, a_j) \text{ s.t. } d(a_i, a_j) = g(L_a, L_b)} \quad (3.25)$$

where L_a and L_b are the two fragments in ligand L , $g(L_a, L_b)$ defines the geometric

constraint between the two fragments and $d(a_i, a_j)$ is the distance between two active site points a_i and a_j . Therefore, the denominator gives the probability of finding any two points across the entire active site that satisfy the geometric constraints (irrespective of the classifications associated with them).

This normalization will account for the ease to find two points that match a given set of distance constraints than it is to find three such points. The normalization procedure can be easily extended to the analysis of ligands with 3 or more fragments and therefore allow for a direct comparison between single-fragment ligands and multiple fragment ligands in a fair manner. This normalization procedure works as follows:

1. Compute the posterior probability of all single fragment ligands based on the feature-based kernel density estimation procedure
2. Compute the joint probability from the *MRF* formulation for all the multi-fragment ligands
3. Apply normalization specified in Equation 3.25 to all the joint probabilities from multi-fragment ligands
4. Pick the ligand with the highest value of $P'(f|L, \Gamma_S)$ across all ligands.

F. Conclusions

In this chapter, a *MRF* approach to the combination of fragments into a multi-fragment ligand has been developed. This probabilistic analysis enables the determination of the correct ligand binding an active site based on the determination of the lowest energy labeling of active site mesh points without enumerating all possible labellings. The *MRF* analysis takes into account the fragment classification based on stereochemical features at each mesh point as well as the geometric constraints

between ligand fragments while finding the ligand most likely to bind any given active site. This analysis provides a formal probabilistic framework to capture the feature information as well as the contextual constraints between ligand fragments.

CHAPTER IV

DATABASE CREATION

In this chapter, Section A describes the creation of a database of ligands which will be used to test the functional analysis algorithms developed throughout this dissertation. This section will introduce a fragment-based analysis of ligands that will limit the effect of ligand flexibility on automated functional analyses. This section will also describe the various ligands in the database, the homology between proteins in each of the fragment classes and finally the fold diversity within each fragment class. Diversity in the database relating to homology and fold families is essential to ensure that the algorithms developed in this dissertation are capable of analyzing the function of apo-proteins even when they have very low sequence homology with the examples in the database and are from diverse fold families. The additional diversity in this database comes from the large number of ligands analyzed. A large-scale automated analysis of protein active sites ensures that the features developed are not specific to certain protein-ligand interactions but capture some basic information about interactions in the active sites.

A. Database Creation and Ligand Classes

The *Protein Data Bank* makes available the structures of many protein-ligand complexes, but, many of these ligands do not necessarily interact with the protein in its active site. In this dissertation, the focus is on catalytically interesting interactions between proteins and ligands. This necessitates that additional information regarding the amino acids that form the active site as well as those that interact with the protein be available for the complexes considered in our databases. *Ez-CatDB* [91] and *Catalytic Site Atlas* [101] are two such databases that have a list

of protein-ligand complexes as well as a list of residues that interact with the ligand. In this study, only complexes where the ligand interacts with the protein in its active site were chosen. A combined list of complexes was created from these two databases and the various ligands in these complexes were identified. Table II lists a sample of the 1217 ligands used to populate the final database. This table also lists the number of examples of each of these ligands. 815 of the 1217 ligands in the database have only one example and 22 of these ligands have 10 or more examples. The ligand with the largest number of examples is *nicotinamide adenine dinucleotide (NAD)* with 117 examples. The complete list of ligands is available at http://saclab.tamu.edu/active_anal/database_ligs.html.

Table II.: A sample of the multi-fragment ligands in the current database

Ligand ID	No. of Examples	No. of Fragments
061	1	5
074	1	3
108	2	2
114	1	3
117	1	5
120	1	2
130	2	3
132	1	3
133	1	3
134	1	3
135	1	2
138	1	6
13P	6	2
146	2	7
155	1	2
157	1	2
166	1	3
16G	2	2
191	1	6
1BO	1	1
1IN	2	6

As mentioned in the introduction (Chapter I), multi-fragment ligands (with greater than 5 rotatable bonds) tend to have greater flexibility in protein-ligand complexes. This flexibility makes it difficult to capture the resultant diversity of the patterns of protein-ligand interactions. The *fragment-based approach* introduced in this dissertation will limit the internal degrees of freedom within each ligand fragment thereby making it easier to use pattern-recognition techniques to capture the interaction patterns. Therefore, each of the 1217 multi-fragment ligands is split into fragments containing no more than 6-7 carbon atoms using *Electronic Ligand Builder and Optimization Workbench (elBOW)* [84]. Some of these ligand fragments are shared by multiple multi-fragment ligands (for example, *adenine* is found in ADP, AMP, ATP, NAD etc.). Additionally, some ligand fragments can be combined into a single class due to their chemical similarity. In this study, the chemical similarity between ligand fragments is based on the similarity between the fingerprints of the ligand fragments that were computed using OpenEye’s implementation of the Tanimoto score [122] defined as follows:

$$Tanimoto(A, B) = \frac{N_c}{N_a + N_b - N_c} \quad (4.1)$$

where N_a and N_b are the number of bits set to 1 in the fingerprint of ligand fragments A and B respectively, and N_c is the number of bits set to 1 in both ligand fragments A and B . The fingerprints themselves are computed using the *makefp* [93] utility. Any two ligands that have a *Tanimoto* score ≥ 0.7 are considered to be similar and placed in the same ligand fragment class. The database has 441 ligand fragment classes, out of which 250 fragment classes have only one example and the class with the maximum number of examples (1434 examples) is *dimethylbutanamide*.

A sample of these fragment classes is listed in Table III. This table also shows the average homology between all the examples within each ligand fragment. In this study, no attempt was made to set any homology cutoff while creating the database and therefore, while most of the examples in each fragment class have an average homology of less than 35% (354/441 classes), there exist some classes with higher average homology values. The proteins in the database additionally also span diverse fold-families (based on the *SCOP* fold classification by [89]). The table also lists the number of fold families represented in each fragment class in the database. The number of SCOP folds in each fragment class ranges from 1 (classes with one example) to 127 (class *dimethylbutanamide* with 1434 examples). Once again, the complete list of fragment classes and the above data for each of the classes is available at http://saclab.tamu.edu/active_anal/database_ligs.html.

Table III.: A sample of the ligand fragment classes in the database and a list of the multi-fragment ligands that contain each of these fragments

Ligand Fragment	Ligands Containing Fragment	Num In Class	Avg % Homol	Num SCOP Folds
dethiobiotin	DTB	2	5.0	2
acetamido aminodihydro pyrancarboxylic acid	49A, 4AM, 936, 9AM, ARH CXN, DAN, DPC, E09, FID G20, G23, G26, G28, G37 GNA, GNT, L34, MCN, NTZ PCD, PN1, R56, TTG, ZMR	43	17.7	12
triphosphate	3AT, 3PO, 4TA, ANP, AP5 ATP, CTP, D3T, DAD, DCP DCT, DG3, DGT, DTP, GP3 GTP, MGT, T5A, TTP, UP5 Z5A	118	8.4	28
methylpiperazine carbaldehyde	1IN, 3IN, BZP, CDX, GEQ MK1, PIN, STI, SU2, UKP	22	19.6	9
pyridoxal	5PA, CBA, DCS, ELP, EPC	158	11.9	9

Table III – Continued

Ligand Fragment	Ligands Containing Fragment	Num In Class	Avg % Homol	Num SCOP Folds
	HCP, HEN, IK2, ILP, IN5 KAM, KET, LCS, MPM, NMA NOP, PDD, PFM, PGU, PLA PLG, PLP, PLS, PLT, PLV PMH, PMP, PP3, PPD, PPE PPG, PY4, PY5, PY6			

Feature-based methodologies are used to analyze the active sites in this study and therefore the diversity in active site patterns are captured with greater accuracy for classes with more examples. While our database contains many classes with only one example, these classes help increase the completeness of the database and also increase the possibility of identifying the function of an unknown test protein.

The final database has 1217 different ligands that can be grouped into 441 fragment classes. The 1217 ligand interaction patterns are analyzed based on 2383 protein-ligand complexes leading to a total of 7070 fragments. This final is very diverse and contains examples belonging to various fold families with mostly low sequence homology to each other. No additional resolution thresholds were applied during database creation (database consists of medium-low resolution structures).

B. Conclusions

This chapter details the 1217 diverse ligand families used as the database during the rest of the dissertation, the homology between examples in each fragment class and also the diverse fold families represented within each class. The proteins in the database were restricted to those for whom information regarding the active site interactions was available in other pre-existing databases. Ongoing efforts to better

annotate structural data in the *PDB* ([52], [53]) will allow for the growth of this database to cover even more ligands.

Almost all existing methodologies are unable to deal with ligand flexibility in a satisfactory manner, thereby reducing the effectiveness of these algorithms. The fragment-based analysis is a novel and an essential improvement to existing functional annotation methods. The creation of the fragment database presented in this chapter is essential to testing the accuracy of the fragment-based methodology presented in this dissertation.

CHAPTER V

RESULTS

In this chapter, the results of the classification and combination methodologies described in Chapter II and III respectively, on the database of ligand fragments described in Chapter IV will be examined. First, in Section B, the need for both geometric and electrostatic features for active site classification will be examined and the analysis will show that including both these sets of features greatly increases the accuracy of fragment classification as opposed to the accuracies seen when using either just geometric or just electrostatic features. In Chapter II, the dimensionality techniques used in this study were presented and in Section D, the effect of these techniques on feature similarity will be analyzed and it will be shown that there is greater similarity between feature vectors belonging to the same fragment class as opposed to those between two different fragment classes after dimensionality reduction. Section E will present the results of the classification methodology on fragment classes with only one example and the results of the classification algorithm for all the other fragment classes using localized as well as position-dependent stereochemical features are presented in Section F. As mentioned in Chapter I, traditional methods of active site analysis depend on similarities in fold families and sequence. The results presented in Section F will also show that both sets of stereochemical features developed in this study capture the interaction patterns that go beyond similarities in sequence and fold families and therefore exclusion of all other examples with high homology or those that belong to the same fold family have minimal effect on the classification accuracy. In Section H, two new metrics for active site analysis that are different from the classification accuracies are presented and the results of these analyses on a subset of the multi-fragment ligands in this study are shown. In Section

I, the results of the fragment combination methodology developed in Chapter III are presented. Finally, in Section J, the results of using this methodology to analyze three hypothetical proteins are presented. For all these test cases, complex structures were unavailable but functional studies based on sequence analysis existed. In all the accuracy tests presented in this chapter, the analysis is based on the mesh representation of the active site and the stereochemical features calculated for the active site. The actual ligands are considered to be absent and only used to define the true class of each of the active sites.

A. Statistical Comparisons of Classification Accuracies

In this study, many parameters have been introduced: geometric versus electrostatic features, localized versus position-dependent features etc. In each case, it is necessary to know the effect of each of these parameters on the accuracy of classification. In machine learning literature, the *paired-t-test* is often used to determine if the accuracies of two different classifiers differ significantly from each other. The *paired-t-test* assumes that the paired differences are independent and identically normally distributed. While, the differences in accuracy in this study are certainly independent, they need not be normally distributed. Therefore, a related test, the *Wilcoxon signed rank test* will be used to compare all pairs of classification schemes. The *Wilcoxon* test computes the signed difference between the two sets of observations (X and Y) and computes the *Wilcoxon* signed rank statistic, W_+ , as follows:

$$W_+ = \sum_{i=1}^n \phi_i R_i \quad (5.1)$$

where R_i is the rank of the ordered Z_i values computed as

$$Z_i = Y_i - X_i \quad \forall i = 1 : n \quad (5.2)$$

and ϕ_i is the sign of the difference given by

$$\phi_i = I(Z_i > 0) \quad (5.3)$$

where $I(Z_i > 0)$ is called an indicator function and is 1 when the condition $Z_i > 0$ is met and 0 otherwise. The z-score of the W_+ statistic is computed as

$$z_w = \frac{W_+ - \mu_w}{\sigma_w} \quad (5.4)$$

where μ_w is the mean of the rank statistic and is assumed to be zero. The null hypothesis or the case when there is no difference between the two sets of observations would yield a W_+ value of 0 and therefore this is chosen as the mean. The σ_w value for a given value of n has been found to be $\frac{n(n+1)}{4}$. Based on this z-score, a *p-value* associated with the statistic is computed and this *p-value* is used to determine the significance of the difference between the two sets of observations X and Y . This test will be used in all the other sections in this chapter to compare the results of all the different classification methodologies. For example, while comparing the accuracy of classification while using localized stereochemical features and the accuracy while using position-dependent features, X_i refers to the classification accuracy of the localized features on class i and Y_i refers to the classification accuracy of the position-dependent features on class i .

B. Both Geometric and Electrostatic Features Are Necessary for Classification

Both sets of features developed in this study (localized and position-dependent features) are used to capture both the geometric and electrostatic variation in the active sites binding the same ligand. The assumption in this design was that active sites that bind the same ligand have similarities between both the geometric and the electro-

static interaction patterns. A preliminary study based on a subset of protein-ligand complexes tested and confirmed this hypothesis using localized stereochemical features [96]. In this study, the database consisted of complexes of 6 different fragment classes, *adenine*, *citrate*, *nicotinamide*, *pyridoxal*, *phosphate* and *ribose*. This database had 55 examples of all classes other than *citrate* and only 16 examples of *citrate*. In this study, all active sites were not created using a uniform radius but the radius of the active site pockets varied based on the size of the fragment being analyzed. In this study, pockets for all fragments other than *phosphate* were drawn at 5Å and the pockets around *phosphate* were created at 4Å

The study explored whether using just geometric features or just the electrostatic features by themselves would yield comparable/better classification accuracies than when using both sets of features. In this study, a total of 37 features: 16 geometric features and 21 electrostatic features were used. After dimensionality reduction using *SVD*, the transformed axes corresponding to the top 6 singular values were chosen for further computations. The *probabilistic kernel density estimation* classifier was used to classify the reduced-dimension feature vectors. Equal prior probabilities were assumed for all ligand classes and the probability density function (described by equation 2.32) can be used to estimate the likelihood of a class C_i given a test feature vector $\Phi(\mathbf{x})$, as $P(C_i | \Phi(\mathbf{x}))$. Each feature vector is classified as belonging to the class with the highest log likelihood.

The results of this analysis are shown in Table IV. The classification accuracy using all the localized stereochemical features greatly increased by over 15% in most cases as compared to either sets of features by themselves. In only one case, using only one set of features had a higher accuracy than using the combined set and that was with the fragment class *phosphate*. In this experiment, as mentioned before, the active site size varied based on the size of the fragment. Since the size of the

phosphate fragment was smaller than all the other fragments, when using geometric features alone, these fragments were classified accurately 100% of the time.

The p-values listed in the table were obtained by comparing the classification accuracy of just geometric and just electrostatic features with all the localized features respectively. The p-values of 0.0042 and 0.0027 show that the differences in the accuracies between these different sets of features are statistically significant. Therefore, this study concluded that there was a significant advantage to using both geometric and electrostatic features when capturing interaction patterns in the active site.

Table IV.: Comparisons of classification accuracy using subsets of localized stereochemical features show that classification accuracy is greatly improved when both geometric and electrostatic features are used to describe the nature of the various active site pockets

Ligand Name	Accuracy Using Both Sets of Features	Accuracy Only Geometric Features	Accuracy Only Electrostatic Features
adenine	38/55 (69%)	25/55 (45%)	29/55 (52%)
citrate	9/16 (56%)	7/16 (44%)	6/16 (38%)
nicotinamide	35/55 (64%)	31/55 (56%)	22/55 (40%)
phosphate	53/55 (96%)	55/55 (100%)	25/55 (45%)
pyridoxal	37/55 (67%)	20/55 (36%)	26/55 (47%)
ribose	35/55 (64%)	24/55 (44%)	23/55 (42%)
p-value		0.0042	0.0027

C. Note Regarding Current Database

Before outlining the accuracies of the algorithms presented in this study, a note regarding the database is essential. There are a total of 441 fragment classes in the database. 403 of these classes have fewer than 10 examples and 240 of these have a single example (a total of 782 examples). These fragment classes have been included

in the database to ensure that the database is general and representative of the complexes currently in the *PDB*. Figure 15 shows the distribution of fragment class sizes in the database for all classes with fewer than 10 examples. Examples from the 38 classes with 10 or more examples are used to determine the most relevant features using the combined *SVD* + *LDA* techniques since the *LDA* projections are determined based on the mean values of feature vectors. This ensures that the computation of mean and standard deviation values were statistically valid. But, the resulting *SVD* and *LDA* projections were used to analyze all classes, a fact that should be taken into account while analyzing the accuracy of the classification algorithms. Additionally, there are no additional fold or sequence homology constraints on members within a fragment class. Therefore, multiple examples within a fragment class could have the same fold family and also have sequence homology greater than 35%.

D. Effect of Dimensionality Reduction Techniques Using a Combination of *SVD* and *LDA* Projections

A combination of *SVD* and *LDA* techniques is used to identify features with the most information and those that help discriminate between various fragment classes. The *SVD* technique is designed to identify and project the data onto the principal directions of variation of the feature vectors and the *LDA* technique projects the data onto directions that best discriminate between the classes. Together, both these techniques help eliminate the noise in feature vectors (information unrelated to classification as well as information unrelated to interaction patterns). The assumption behind using these techniques is that the feature similarity between feature vectors belonging to the same fragment class will be increased after these techniques are applied as opposed to when using the raw feature vectors.

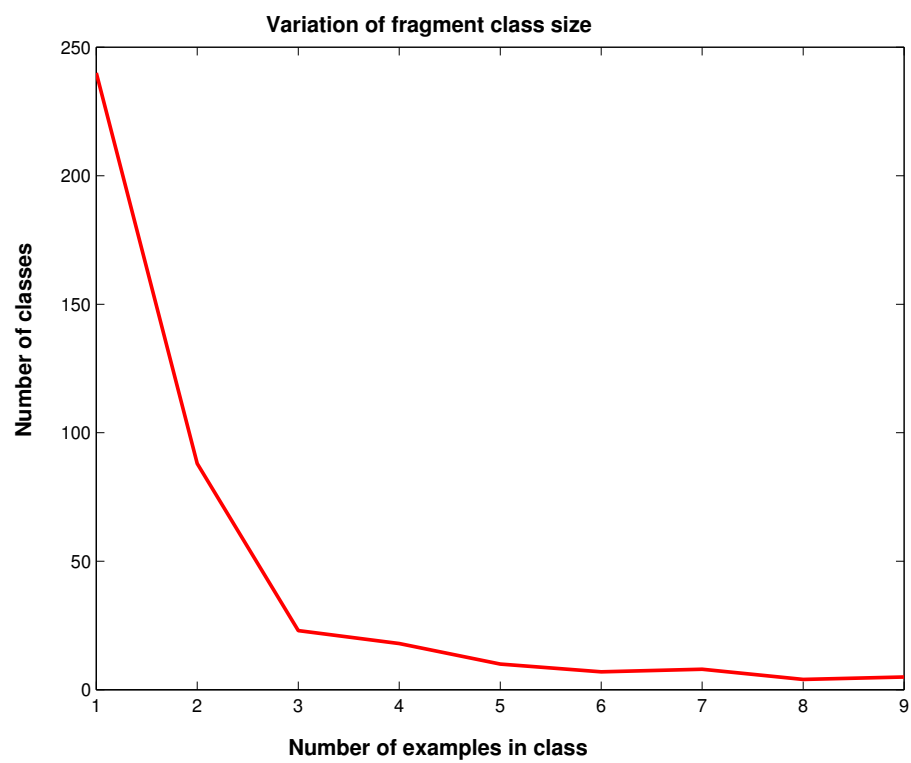


Fig. 15. Distribution of number of examples in various fragment classes with fewer than 10 examples

For 10 randomly selected fragment classes the following statistic was computed:

$$R = \frac{\text{avg. feature difference between classes}}{\text{avg. feature difference within class}} \quad (5.5)$$

Figure 16 shows the variation of R for 10 randomly selected fragment classes. These ratios were computed using all the raw feature vectors, using only the *SVD* projections and finally using both the *SVD* and *LDA* projections. It is clear that using the combination technique of *SVD* and *LDA* techniques greatly improves the discrimination between feature vectors that belong to the same class and those that do not belong to the same class. The use of the combined technique of *SVD* and *LDA* improves the discrimination between classes very clearly in some classes like class 4 and class 6 in Figure 16 while for other classes like class 7 and class 9, there is no significant improvement. This shows that while the principal directions of data variance (as determined by *SVD*) are sufficient to separate out information between some classes, in some cases better separation between classes is obtained by using the *LDA* projections.

E. Fragment Classification Analysis for Classes with One Example

For the 240 fragment classes with one single example, it is not possible to find another closest example of the same class and therefore not possible to classify it correctly. The next best thing would be to examine if the classification scheme consistently picked examples from the database most similar to these test cases thus ensuring that the classification scheme was recognizing similarities in interaction patterns even with only one example. The similarity between the test example and those picked by the classification schemes was determined based on the *Tanimoto* similarity. Therefore, for each of the 240 single examples, the *Tanimoto* score to their closest example in the

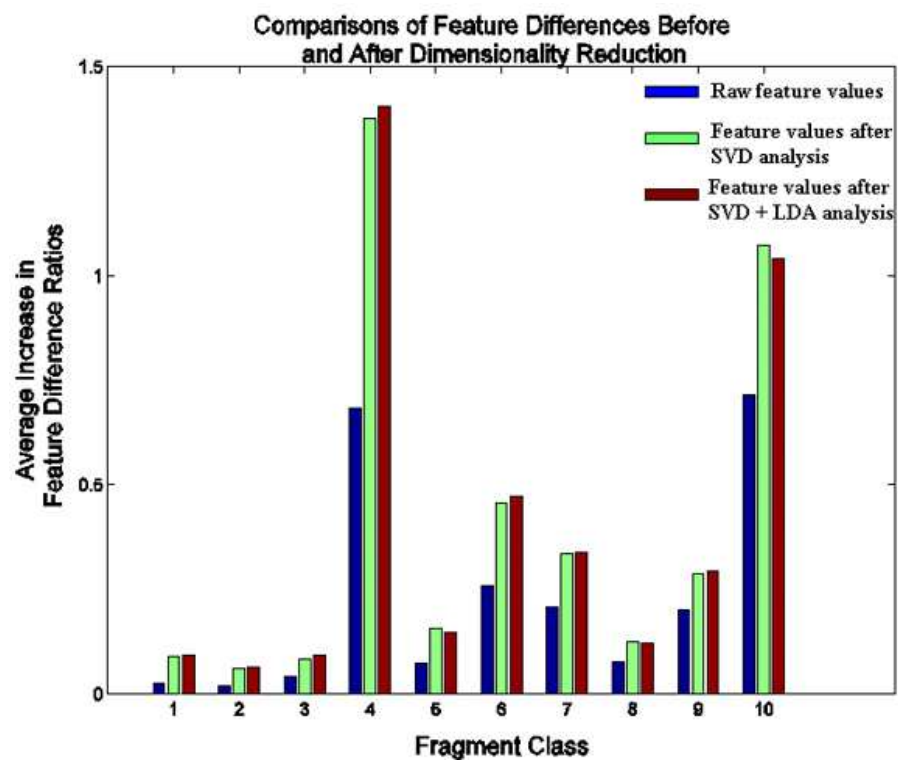


Fig. 16. A combination of *SVD* and *LDA* techniques have enabled the selection of features with the most relevant information regarding active site interaction patterns

entire database was found and this score was compared to the highest *Tanimoto* score based on the classification scheme using position-dependent stereochemical features. Figure 17 shows the overall closest example to the test fragment (shown in Figure 17(a)) as well as the closest match found during classification and Figure 18 shows an example where the overall closest in the database was also the closest match found during classification for the test fragment in Figure 18(a).

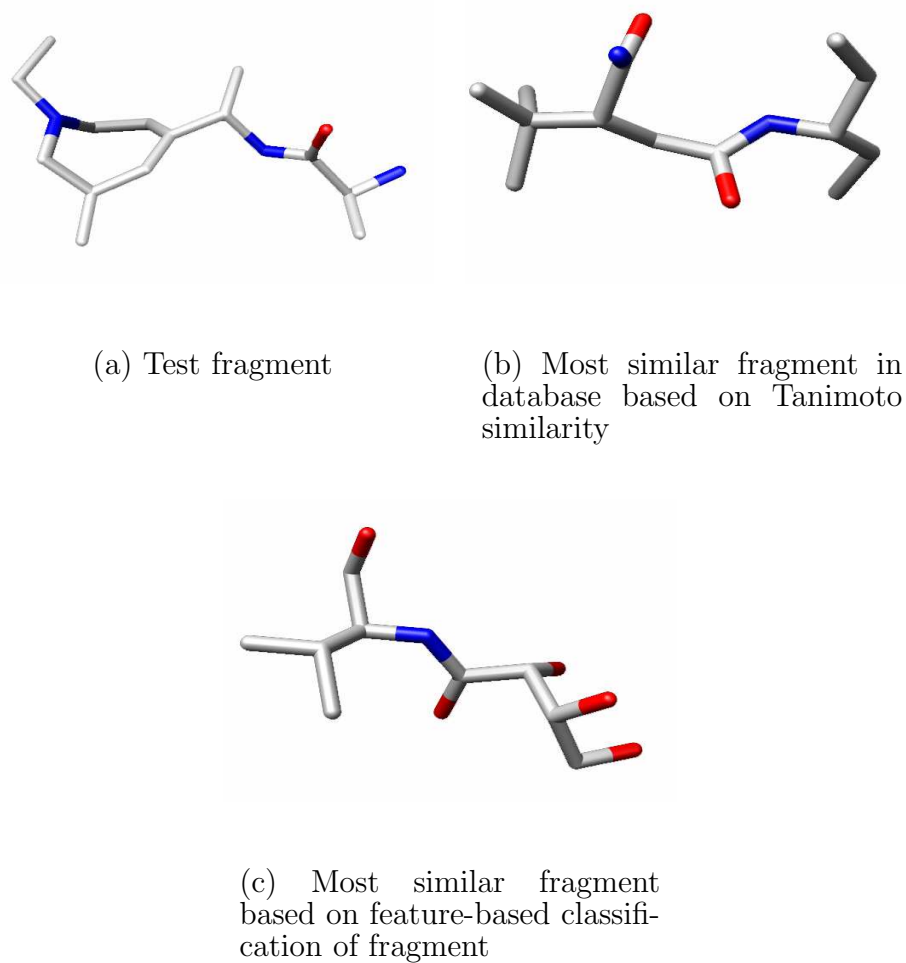


Fig. 17. Feature-based classification yields a match very similar to the fragment with the highest Tanimoto similarity for a test fragment from a single fragment class

Figure 19 shows this comparison for all the 240 single fragment classes. This figure shows the correlation between the *Tanimoto* similarity of each single fragment example with the closest fragment in the database based on chemical similarity, X , and the *Tanimoto* similarity of each single fragment example with its closest match based on the stereochemical feature vectors, Y . The correlation coefficient between these values was computed as

$$r_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{(n-1) s_x s_y} \quad (5.6)$$

where \bar{x} is the mean of all the values in X , \bar{y} is the mean of all the values in Y , s_x is the standard deviation of all the values in X and finally, s_y is the standard deviation of all the values in Y and n is the number of measurements in X and Y . The correlation coefficient was found to be 0.22. The significance of the correlation coefficient was computed by looking up the value of t in a table for a given df value, where $df = n - 2$. This significance value was found to be 0.0003 which shows that a correlation coefficient of 0.22 for the 240 *Tanimoto* similarity values is significant. While the correlation coefficient is not very high, its significance is encouraging since the *Tanimoto* similarity is purely based on chemical similarity and therefore in cases where the *Tanimoto* similarity is not as high for matches from the classification scheme, the geometric features could very well have driven the feature matching process.

F. Analysis of Classification

In this section, the accuracy of classification using both the localized stereochemical features as well as the position-dependent stereochemical features will be examined. These features are calculated at mesh points closest to the center of each of the fragments in the database. The position-dependent features capture the interaction

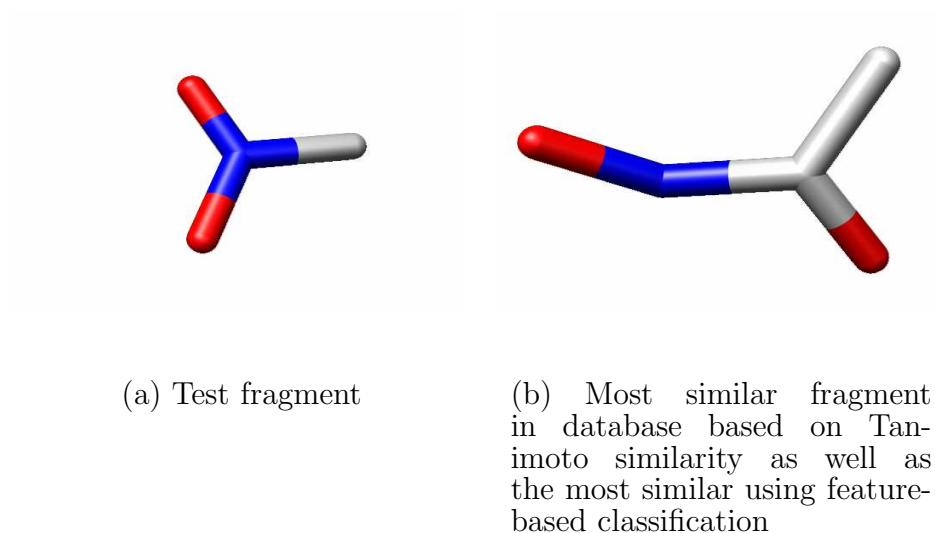


Fig. 18. Match based on feature-based classification is the same as the one with the highest Tanimoto similarity for a test fragment from a single fragment class

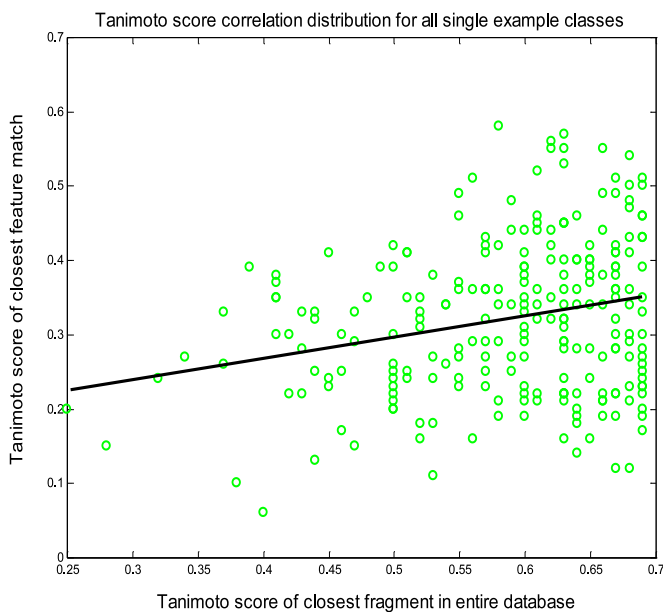


Fig. 19. Comparison of Tanimoto scores for classes with one example

patterns more accurately since the basis for these features is a canonical representation of the active sites. The canonical representations enables a more detailed specification of interaction patterns, for *e.g.*, it is possible to look at specific regions of the active site based on 3D coordinates and compare patterns between active sites. In this section, the accuracy of the classification scheme presented in Chapter II using both these sets of features will be tabulated. In the analysis of both sets of features, the classes with K highest probability values for each example in the database are computed and an example is said to be correctly classified if the actual fragment class for the example exists within these K classes. In this study, the value of K was determined to be 10 based on an analysis of the accuracy of the classification scheme on the database for various K values. Figure 20 shows the results of this analysis.

Based on this definition of accuracy, Table V details the accuracy of using localized stereochemical features to classify each patch in the database with a fragment class. This table again confirms that using both geometric and electrostatic features yields higher classification accuracy than using either only geometric or only electrostatic features (77.6%, 73.6% and 73.0% respectively). The table also lists the p-values (at the bottom of the table) computed as described in Section A and the p-values show that the differences in accuracy between using all stereochemical features and just the geometric as well as just the electrostatic are both statistically very significant. This underlines the importance of capturing stereochemical patterns within active sites. This table lists 58 fragment classes (30 with 10 or more examples) with one more example correctly classified using either geometric, electrostatic or the combined set of features.

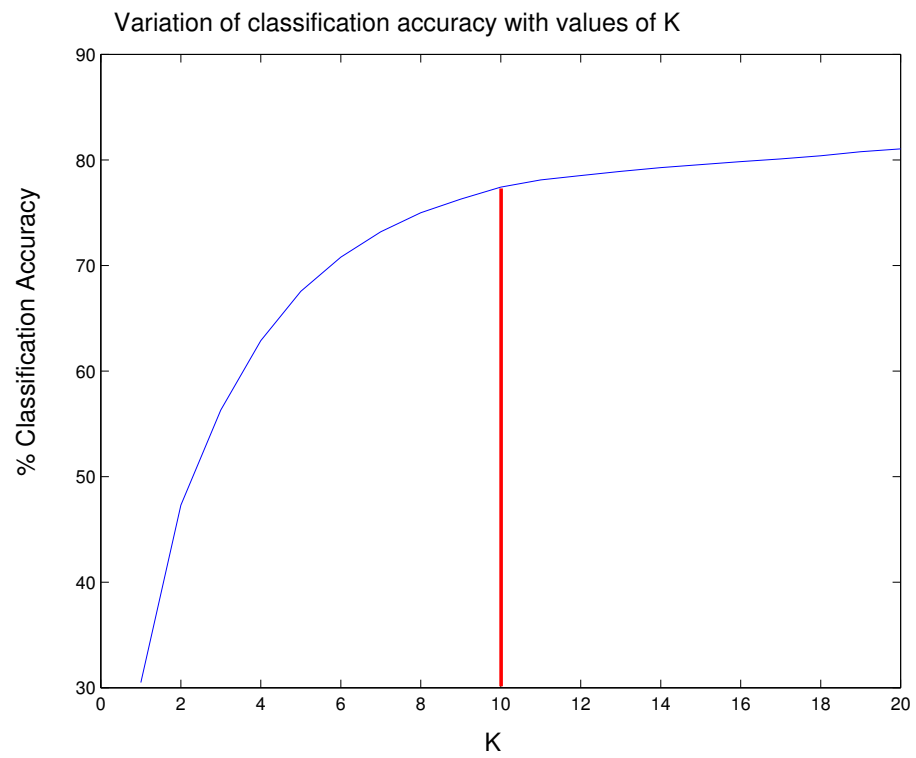


Fig. 20. Analysis of database accuracy for various K values

Table V.: Comparison of using localized stereochemical features for fragment classification

Ligand Fragment	Number Examples	Accuracy Geometric Features	Accuracy Electrostatic Features	Accuracy Combined Features
phosphate	671	581 (86.6%)	544 (81.1%)	614 (91.5%)
dimethylbutanamide	1436	1384 (96.4%)	1397 (97.3%)	1423 (99.1%)
diphosphate	545	440 (80.7%)	429 (78.7%)	489 (89.7%)
acetamido aminodihydro				
pyrancarboxylic acid	43	9 (20.9%)	17 (39.5%)	15 (34.9%)
triphosphate	120	36 (30.0%)	63 (52.5%)	58 (48.3%)
adenine	699	617 (88.3%)	599 (85.7%)	666 (95.3%)
methylpiperazine				
carbaldehyde	22	3 (13.6%)	1 (4.6%)	3 (13.6%)
pyridoxal	158	104 (65.8%)	115 (72.8%)	125 (79.1%)
dihydroxy tetramethyl				
diazepanone	18	14 (77.8%)	11 (61.1%)	15 (83.3%)
dichlorophenol	12	3 (25.0%)	4 (33.3%)	3 (25.0%)
methylchromanone	13	0 (0.0%)	1 (7.7%)	0 (0.0%)
ribose	1329	1297 (97.6%)	1299 (97.7%)	1327 (99.9%)
sulfuric acid	37	1 (2.7%)	2 (5.4%)	0 (0.0%)
p-tolylcarbinol	16	3 (18.8%)	5 (31.3%)	3 (18.8%)
thymine	163	91 (55.8%)	71 (43.6%)	93 (57.1%)
ascorbic acid	4	2 (50.0%)	0 (0.0%)	0 (0.0%)
nicotinamide	223	116 (52.0%)	110 (49.3%)	123 (55.2%)
methylhypoxanthine	5	1 (20.0%)	4 (80.0%)	3 (60.0%)
normotiroide	10	0 (0.0%)	1 (10.0%)	0 (0.0%)
diisopropyl hydrogen				
phosphite	9	0 (0.0%)	0 (0.0%)	1 (11.1%)
nitrobenzene	5	0 (0.0%)	1 (20.0%)	0 (0.0%)
butyl dimethylindole	50	6 (12.0%)	19 (38.0%)	12 (24.0%)
pantothenamide	12	2 (16.7%)	4 (33.3%)	2 (16.7%)
azidoribose	7	1 (14.3%)	3 (42.9%)	0 (0.0%)
androstanedione	25	1 (4.0%)	0 (0.0%)	0 (0.0%)
aminomethyl cyclopentanol	4	0 (0.0%)	4 (100.0%)	3 (75.0%)
diaminomethylidene				
aminoethylpentanyl				
carboxylic acid	3	0 (0.0%)	1 (33.3%)	0 (0.0%)
caprylene	19	1 (5.3%)	4 (21.1%)	2 (10.5%)
ethylsulfanyl				
isocyanatoethane	4	0 (0.0%)	1 (25.0%)	0 (0.0%)
picoline	6	0 (0.0%)	2 (33.3%)	2 (33.3%)
hemineurine	7	1 (14.3%)	5 (71.4%)	4 (57.1%)

Table V – Continued

Ligand Fragment	Number Examples	Accuracy Geometric Features	Accuracy Electrostatic Features	Accuracy Combined Features
beta arabino				
furanosylamine	79	13 (16.5%)	12 (15.2%)	10 (12.7%)
iodopyrazole	6	2 (33.3%)	3 (50.0%)	1 (16.7%)
benzothiophene				
carboximidamide	3	2 (66.7%)	0 (0.0%)	1 (33.3%)
isoequilenin	16	10 (62.5%)	2 (12.5%)	6 (37.5%)
aminomethyl methylpyrrolo				
pyrimidinone	6	0 (0.0%)	2 (33.3%)	0 (0.0%)
amino hydroxyindane	15	6 (40.0%)	13 (86.7%)	13 (86.7%)
fluoromethylbenzene	11	2 (18.2%)	2 (18.2%)	3 (27.3%)
butyl alcohol	212	125 (59.0%)	88 (41.5%)	116 (54.7%)
dimethylimidazole	11	1 (9.1%)	0 (0.0%)	0 (0.0%)
cyclopentanamine	14	0 (0.0%)	1 (7.1%)	0 (0.0%)
hydroxyethylmethyl				
imidazole carboxamide	9	3 (33.3%)	7 (77.8%)	5 (55.6%)
dihydroxypyrrolidin				
ethanone	5	1 (20.0%)	0 (0.0%)	0 (0.0%)
dichloropiperazine	3	2 (66.7%)	3 (100.0%)	2 (66.7%)
methyltrifluoromethyl				
pyrimidinone	14	1 (7.1%)	5 (35.7%)	2 (14.3%)
nitrophenol	7	1 (14.3%)	0 (0.0%)	0 (0.0%)
ethyl methylthiazole	9	3 (33.3%)	1 (11.1%)	1 (11.1%)
aminoethyl benzene				
sulfonic acid	19	1 (5.3%)	1 (5.3%)	0 (0.0%)
diamino quinazolinone	13	4 (30.8%)	8 (61.5%)	5 (38.5%)
diiodo methylphenol	6	0 (0.0%)	1 (16.7%)	0 (0.0%)
chloromethylbenzene	10	0 (0.0%)	1 (10.0%)	0 (0.0%)
acetamido methyl				
boronic acid	6	0 (0.0%)	1 (16.7%)	1 (16.7%)
diaminopenetenoic acid	4	0 (0.0%)	1 (25.0%)	0 (0.0%)
flavin	131	99 (75.6%)	52 (39.7%)	92 (70.2%)
aminomethyl benzimidazolyl				
methylidene azanium	63	21 (33.3%)	44 (69.8%)	38 (60.3%)
methionine	26	6 (23.1%)	10 (38.5%)	8 (30.8%)
thiamin	9	8 (88.9%)	7 (77.8%)	8 (88.9%)
cobinamide dihydrate	7	3 (42.9%)	3 (42.9%)	3 (42.9%)
Total	6830	5028 (73.6%)	4985 (73.0%)	5301 (77.6%)
p-value		0.0071	0.0016	

The top fragment class matches for a *nicotinamide* fragment and the corresponding probabilities are listed in Table VI. This table shows that the *nicotinamide* fragment is accurately identified with a relatively high posterior probability of 0.3. Additionally, the next closest fragment class is *adenine* which is a very similar fragment to *nicotinamide* showing that the top matches returned by the classification scheme do indeed capture the stereochemical interaction patterns within the active site.

Table VII shows a similar analysis using positional-dependent stereochemical features. This table again confirms that using both geometric and electrostatic features yields higher classification accuracy than using either only geometric or only electrostatic features (84.2%, 77.8% and 81.5% respectively). Once again, the p-values are computed and listed in the table and the results show that the difference in accuracy between using all the stereochemical features and using only geometric features is extremely statistically significant and the difference in accuracy between using all the stereochemical features and using only electrostatic features is significant. This table lists 76 fragment classes (35 out of 76 with greater than 10 examples) with one more example correctly classified using either geometric, electrostatic or the combined set of features.

Table VI.: Top 10 matches and corresponding probabilities for a *nicotinamide* fragment

Fragment Name	Probability from <i>KDE</i>
nicotinamide	0.3
adenine	0.2
dimethylbutanamide	0.1
pyridoxal	0.09
butyl alcohol	0.05
phosphate	0.03
diphosphate	0.01

Table VI – Continued

Fragment Name	Probability from <i>KDE</i>
triphosphate	0.01
cyclopentanamine	0.01
iminoglycine	0.01

There is a 7% increase in accuracy using position-dependent features as opposed to the localized stereochemical features which corresponds to a statistically significant p-value of < 0.0001 . Additionally, the number of fragment classes with one or more example correctly classified has increased from 58 to 76 (out of a possible 200 fragment classes with greater than one example in the database). Both these factors indicate that the position-dependent features are more capable of capturing interaction patterns more accurately and also for a larger number of fragment classes.

Previous feature-based methodologies were only able to distinguish between binding sites and non-binding sites with accuracy around 60% [46]. One of the claims in this study was that global features would not be sufficient to characterize interaction patterns and additionally distinguish between patterns across different fragment class. This claim has now been affirmed since there is a 24% increase in classification accuracy between the use of global features described in [46] and the use of position-dependent stereochemical features for active site analysis. Additionally, this increased accuracy is obtained in the analysis of 441 different fragment classes (as opposed to the two-class problem solved earlier). Since the position-dependent features provide the most-detailed view of the active site stereochemistry, the increased accuracies using these features is further evidence that active site interaction patterns are captured with greater accuracy by increasing the granularity of the stereochemical features used to capture these patterns.

Table VII.: Results of using geometric and electrostatic position-dependent features for classification

Ligand Fragment	Number Examples	Accuracy Geometric Features	Accuracy Electrostatic Features	Accuracy Combined Features
phosphate	671	637 (94.9%)	642 (95.7%)	665 (99.1%)
dimethylbutanamide	1436	1433 (99.8%)	1414 (98.5%)	1435 (99.9%)
diphosphate	545	440 (80.7%)	489 (89.7%)	522 (95.8%)
acetamido aminodihydro				
pyrancarboxylic acid	43	14 (32.6%)	14 (32.6%)	17 (39.5%)
triphosphate	120	29 (24.2%)	63 (52.5%)	63 (52.5%)
adenine	699	682 (97.6%)	655 (93.7%)	697 (99.7%)
methylpiperazine				
carbaldehyde	22	4 (18.2%)	5 (22.7%)	6 (27.3%)
pyridoxal	158	118 (74.7%)	134 (84.8%)	142 (89.9%)
dihydroxy tetramethyl				
diazepanone	18	9 (50.0%)	11 (61.1%)	12 (66.7%)
ribose	1329	1315 (98.9%)	1306 (98.3%)	1326 (99.8%)
sulfuric acid	37	3 (8.1%)	4 (10.8%)	8 (21.6%)
p-tolylcarbinol	16	8 (50.0%)	8 (50.0%)	10 (62.5%)
thymine	163	104 (63.8%)	118 (72.4%)	129 (79.1%)
ascorbic acid	4	2 (50.0%)	1 (25.0%)	2 (50.0%)
nicotinamide	223	100 (44.8%)	179 (80.3%)	187 (83.9%)
hydroxy methylamino				
pyridazin ethanone	3	0 (0.0%)	1 (33.3%)	2 (66.7%)
thieno pyridine				
carboximidamide	2	1 (50.0%)	0 (0.0%)	2 (100.0%)
aminoethyl carboxypropyl				
phosphoryl	3	0 (0.0%)	0 (0.0%)	1 (33.3%)
propylimidazole	5	0 (0.0%)	1 (20.0%)	1 (20.0%)
normotiroid	10	7 (70.0%)	7 (70.0%)	7 (70.0%)
nitrobenzene	5	1 (20.0%)	0 (0.0%)	1 (20.0%)
butyl dimethylindole	50	13 (26.0%)	16 (32.0%)	21 (42.0%)
pantothenamide	12	1 (8.3%)	0 (0.0%)	0 (0.0%)
trihydroxycyclohexene				
carboxylic acid	5	0 (0.0%)	1 (20.0%)	1 (20.0%)
androstanedione	25	0 (0.0%)	0 (0.0%)	1 (4.0%)
aminomethyl				
cyclopentanol	4	1 (25.0%)	0 (0.0%)	2 (50.0%)
benzene	11	2 (18.2%)	5 (45.5%)	5 (45.5%)
caprylene	19	1 (5.3%)	3 (15.8%)	5 (26.3%)
aminopyrazine				
carbaldehyde	2	1 (50.0%)	0 (0.0%)	2 (100.0%)

Table VII – Continued

Ligand Fragment	Number Examples	Accuracy Geometric Features	Accuracy Electrostatic Features	Accuracy Combined Features
valienamine	7	1 (14.3%)	1 (14.3%)	1 (14.3%)
ethylsulfanyl				
isocyanatoethane	4	1 (25.0%)	0 (0.0%)	1 (25.0%)
difluorobenzyl				
alcohol	5	0 (0.0%)	2 (40.0%)	2 (40.0%)
picoline	6	1 (16.7%)	1 (16.7%)	2 (33.3%)
trihydroxy methyl				
aminohexanal	4	0 (0.0%)	0 (0.0%)	1 (25.0%)
hemineurine	7	0 (0.0%)	1 (14.3%)	0 (0.0%)
beta arabino				
furanosylamine	79	20 (25.3%)	33 (41.8%)	36 (45.6%)
iodopyrazole	6	0 (0.0%)	2 (33.3%)	2 (33.3%)
isoequilenin	16	4 (25.0%)	4 (25.0%)	10 (62.5%)
aminomethyl methylpyrrolo				
pyrimidinone	6	2 (33.3%)	0 (0.0%)	1 (16.7%)
amino difluorohydroxy				
methylheptanal	7	0 (0.0%)	3 (42.8%)	2 (28.6%)
amino hydroxyindane	15	10 (66.7%)	8 (53.3%)	10 (66.7%)
fluoromethylbenzene	11	2 (18.2%)	3 (27.3%)	2 (18.2%)
aminomethyl pyrimidine	3	0 (0.0%)	3 (100.0%)	2 (66.7%)
methylpyridinone	3	1 (33.3%)	0 (0.0%)	1 (33.3%)
271	3	2 (66.7%)	2 (66.7%)	2 (66.7%)
sulfate	12	1 (8.3%)	0 (0.0%)	0 (0.0%)
pentanimidamide	4	0 (0.0%)	1 (25.0%)	0 (0.0%)
dioxothiadiazepane	2	0 (0.0%)	2 (100.0%)	0 (0.0%)
ethanimidoyl				
piperidinol	5	2 (40.0%)	3 (60.0%)	3 (60.0%)
tertbutyl ethyl				
carbamate	8	0 (0.0%)	2 (25.0%)	1 (12.5%)
butane	9	2 (22.2%)	1 (11.1%)	2 (22.2%)
butyl alcohol	212	135 (63.7%)	174 (82.1%)	166 (78.3%)
formanilide	3	2 (66.7%)	0 (0.0%)	0 (0.0%)
dimethylimidazole	11	0 (0.0%)	3 (27.3%)	0 (0.0%)
cyclopentanamine	14	1 (7.1%)	2 (14.3%)	1 (7.1%)
methylpyrrolidinol	3	3 (100.0%)	2 (66.7%)	3 (100.0%)
hydroxyethylmethyl				
imidazole carboxamide	9	7 (77.8%)	7 (77.8%)	7 (77.8%)
dihydroxypyrrolidin				
ethanone	5	0 (0.0%)	1 (20.0%)	0 (0.0%)
guanidinobutanal	8	0 (0.0%)	2 (25.0%)	0 (0.0%)

Table VII – Continued

Ligand Fragment	Number Examples	Accuracy Geometric Features	Accuracy Electrostatic Features	Accuracy Combined Features
diethylbenzo thiophene	4	0 (0.0%)	1 (25.0%)	0 (0.0%)
ethylenylpropanamide	2	0 (0.0%)	2 (100.0%)	0 (0.0%)
methyltrifluoromethyl				
pyrimidinone	14	8 (57.1%)	8 (57.1%)	8 (57.1%)
amino methyl				
pyridinone	2	0 (0.0%)	2 (100.0%)	1 (50.0%)
ethanimidoylecyclo				
hexanamine	7	0 (0.0%)	2 (28.6%)	1 (14.3%)
ethyl methylthiazole	9	3 (33.3%)	3 (33.3%)	3 (33.3%)
hydroxymethyl				
phenylpentanamide	3	0 (0.0%)	2 (66.7%)	0 (0.0%)
benzyl methanoate	4	0 (0.0%)	1 (25.0%)	0 (0.0%)
aminoethyl benzene				
sulfonic acid	19	7 (36.8%)	5 (26.3%)	5 (26.3%)
diamino quinazolinone	13	2 (15.4%)	3 (23.1%)	2 (15.4%)
chloromethylbenzene	10	0 (0.0%)	2 (20.0%)	0 (0.0%)
acetamido methyl				
boronic acid	6	4 (66.7%)	4 (66.7%)	4 (66.7%)
flavin	131	109 (83.2%)	117 (89.3%)	122 (93.1%)
aminomethyl benzimidazolyl				
methylidene azanium	63	41 (65.1%)	47 (74.6%)	46 (73.0%)
methionine	26	7 (26.9%)	16 (61.5%)	17 (65.4%)
thiamin	9	5 (55.6%)	5 (55.6%)	5 (55.6%)
cobinamide dihydrate	7	3 (42.9%)	6 (85.7%)	4 (57.1%)
Total	6830	5312 (77.8%)	5566 (81.5%)	5748 (84.2%)
p-value		< 0.0001	0.0033	

G. Fold Family and Homology Analysis

The aim of this methodology is to use position-dependent features that go beyond specific residues in specific positions to capture the diversity between active sites that bind the same ligand fragment while at the same time discriminating between active sites that bind different ligands. Previous methodologies based active site similarity analysis on the similarities between protein fold families and sequence homology. In

order to ensure that the similarities captured by this methodology are not due to fold similarity, the analysis of each test case was performed by eliminating all training examples that belong to the same fold family. This rigorous test is designed to show that this methodology will be able to capture similarities between examples that bind similar ligands but have little/no fold family similarities.

Similarly, another rigorous test was to employ a sequence homology constraint. While evaluating a test case, all training examples that had sequence homology of 35% or greater were ignored during feature-matching and the test case was compared against all the other examples in the database.

Table VIII shows the results of eliminating members of the same fold family and the results of eliminating homologous sequences during classification using localized stereochemical features. This table shows that there is a greater drop in accuracy when members of the same fold family are ignored during feature comparisons than when ignoring homologous sequences (9% decrease and 2% decrease in accuracy respectively). The table also lists the p-values of comparisons of accuracy after eliminating members of the same fold family and the results of eliminating homologous sequences as < 0.0001 (extremely statistically significant) and 0.0019 (statistically significant) respectively.

Table VIII.: Comparison of leaving fold family out and leaving out homologous sequences during classification using localized stereochemical features

Ligand Fragment	Number Examples	Accuracy Combined Features	Leave Fold Family Out	Remove Homologous Sequences
phosphate	671	614 (91.5%)	545 (81.2%)	592 (88.2%)
dimethylbutanamide	1436	1423 (99.0%)	1351 (94.0%)	1378 (96.0%)
diphosphate	545	489 (89.7%)	374 (68.6%)	452 (82.9%)
acetamido aminodihydro pyranicarboxylic acid	43	15 (34.9%)	1 (2.3%)	13 (30.2%)

Table VIII – Continued

Ligand Fragment	Number Examples	Accuracy Combined Features	Leave Fold Family Out	Remove Homologous Sequences
triphosphate	120	58 (48.3%)	63 (52.5%)	69 (57.5%)
adenine	699	666 (95.3%)	610 (87.3%)	628 (89.8%)
methylpiperazine				
carbaldehyde	22	3 (13.6%)	0 (0.0%)	4 (18.2%)
pyridoxal	158	125 (79.1%)	64 (40.5%)	122 (77.2%)
dihydroxy tetramethyl				
diazepanone	18	15 (83.3%)	0 (0.0%)	9 (50.0%)
dichlorophenol	12	3 (25.0%)	0 (0.0%)	2 (16.7%)
ribose	1329	1327 (99.9%)	1294 (97.4%)	1307 (98.3%)
sulfuric acid	37	0 (0.0%)	0 (0.0%)	1 (2.7%)
p-tolylcarbinol	16	3 (18.8%)	0 (0.0%)	4 (25.0%)
thymine	163	93 (57.1%)	54 (33.1%)	91 (55.8%)
nicotinamide	223	123 (55.2%)	88 (39.5%)	120 (53.8%)
methylhypoxanthine	5	3 (60.0%)	0 (0.0%)	0 (0.0%)
propylimidazole	5	0 (0.0%)	0 (0.0%)	2 (40.0%)
diisopropyl hydrogen				
phosphite	9	1 (11.1%)	0 (0.0%)	1 (11.1%)
butyl dimethylindole	50	12 (24.0%)	0 (0.0%)	14 (28.0%)
pantothenamide	12	2 (16.7%)	0 (0.0%)	2 (16.7%)
trihydroxycyclohexene				
carboxylic acid	5	0 (0.0%)	0 (0.0%)	1 (20.0%)
androstanedione	25	0 (0.0%)	0 (0.0%)	1 (4.0%)
aminomethyl cyclopentanol	4	3 (75.0%)	0 (0.0%)	0 (0.0%)
benzene	11	0 (0.0%)	0 (0.0%)	1 (9.1%)
caprylene	19	2 (10.5%)	0 (0.0%)	1 (5.3%)
ethylsulfanyl				
isocyanatoethane	4	0 (0.0%)	1 (25.0%)	3 (75.0%)
picoline	6	2 (33.3%)	0 (0.0%)	0 (0.0%)
hemineurine	7	4 (57.1%)	0 (0.0%)	1 (14.3%)
beta arabino				
furanosylamine	79	10 (12.7%)	4 (5.1%)	12 (15.2%)
iodopyrazole	6	1 (16.7%)	2 (33.3%)	2 (33.3%)
benzothiophene				
carboximidamide	3	1 (33.3%)	0 (0.0%)	0 (0.0%)
isoequilenin	16	6 (37.5%)	5 (31.2%)	12 (75.0%)
amino hydroxyindane	15	13 (86.7%)	0 (0.0%)	5 (33.3%)
fluoromethylbenzene	11	3 (27.3%)	0 (0.0%)	1 (9.1%)
butyl alcohol	212	116 (54.7%)	106 (50.0%)	122 (57.6%)
cyclopentanamine	14	0 (0.0%)	0 (0.0%)	1 (7.1%)
hydroxyethylmethyl				

Table VIII – Continued

Ligand Fragment	Number Examples	Accuracy Combined Features	Leave Fold Family Out	Remove Homologous Sequences
imidazole carboxamide	9	5 (55.6%)	0 (0.0%)	2 (22.2%)
dichloropiperazine	3	2 (66.7%)	0 (0.0%)	0 (0.0%)
methyltrifluoromethyl				
pyrimidinone	14	2 (14.3%)	0 (0.0%)	2 (14.3%)
nitrophenol	7	0 (0.0%)	2 (28.6%)	2 (28.6%)
ethanimidoylcyclo				
hexanamine	7	0 (0.0%)	0 (0.0%)	1 (14.3%)
ethyl methylthiazole	9	1 (11.1%)	1 (11.1%)	6 (66.7%)
diamino quinazolinone	13	5 (38.5%)	0 (0.0%)	5 (38.5%)
acetamido methyl				
boronic acid	6	1 (16.7%)	0 (0.0%)	0 (0.0%)
flavin	131	92 (70.2%)	95 (72.5%)	109 (83.2%)
aminomethyl benzimidazolyl				
methylidene azanium	63	38 (60.3%)	0 (0.0%)	28 (44.4%)
methionine	26	8 (30.8%)	2 (7.7%)	11 (42.3%)
thiamin	9	8 (88.9%)	0 (0.0%)	7 (77.8%)
cobinamide dihydrate	7	3 (42.9%)	1 (14.3%)	1 (14.3%)
Total	6830	5301 (77.6%)	4663 (68.3%)	5148 (75.3%)
p-value			< 0.0001	0.0019

Similarly, Table IX shows the results of the fold family analysis and the sequence homology analysis based on position-dependent stereochemical features. The table shows that there is a decrease of 6% while ignoring members of the same fold family while there is less than 1% decrease in accuracy when ignoring homologous sequences. The table also lists the p-values of comparisons of accuracy after eliminating members of the same fold family and the results of eliminating homologous sequences as < 0.0001 (extremely statistically significant) and 0.0002 (statistically significant) respectively.

Table IX.: Comparison of leaving fold family out and leaving out homologous sequences during classification using position-dependent stereochemical features

Ligand Fragment	Number Examples	Accuracy Combined Features	Leave Fold Family Out	Remove Homologous Sequences
phosphate	671	665 (99.1%)	663 (98.8%)	665 (99.1%)
dimethylbutanamide	1436	1435 (99.9%)	1434 (99.9%)	1435 (99.9%)
diphosphate	545	522 (95.8%)	510 (93.6%)	523 (96.0%)
acetamido aminodihydro				
pyrancarboxylic acid	43	17 (39.5%)	2 (4.6%)	13 (30.2%)
triphosphate	120	63 (52.5%)	46 (38.3%)	61 (50.8%)
adenine	699	697 (99.7%)	693 (99.1%)	697 (99.7%)
methylpiperazine				
carbaldehyde	22	6 (27.3%)	0 (0.0%)	3 (13.6%)
pyridoxal	158	142 (89.9%)	90 (57.0%)	141 (89.2%)
dihydroxy tetramethyl				
diazepanone	18	12 (66.7%)	0 (0.0%)	9 (50.0%)
ribose	1329	1326 (99.8%)	1324 (99.6%)	1326 (99.8%)
sulfuric acid	37	8 (21.6%)	5 (13.5%)	7 (18.9%)
p-tolylcarbinol	16	10 (62.5%)	0 (0.0%)	8 (50.0%)
thymine	163	129 (79.1%)	88 (54.0%)	127 (77.9%)
ascorbic acid	4	2 (50.0%)	0 (0.0%)	3 (75.0%)
nicotinamide	223	187 (83.9%)	163 (73.1%)	185 (83.0%)
hydroxy methylamino				
pyridazin ethanone	3	2 (66.7%)	0 (0.0%)	2 (66.7%)
aminomethyl imidazol	2	0 (0.0%)	0 (0.0%)	2 (100.0%)
thieno pyridine				
carboximidamide	2	2 (100.0%)	0 (0.0%)	2 (100.0%)
aminoethyl carboxypropyl				
phosphoryl	3	1 (33.3%)	0 (0.0%)	1 (33.3%)
propylimidazole	5	1 (20.0%)	0 (0.0%)	1 (20.0%)
normotiroid	10	7 (70.0%)	0 (0.0%)	6 (60.0%)
nitrobenzene	5	1 (20.0%)	0 (0.0%)	1 (20.0%)
butyl dimethylindole	50	21 (42.0%)	11 (22.0%)	20 (40.0%)
trihydroxycyclohexene				
carboxylic acid	5	1 (20.0%)	0 (0.0%)	1 (20.0%)
androstanedione	25	1 (4.0%)	1 (4.0%)	1 (4.0%)
aminomethyl cyclopentanol	4	2 (50.0%)	0 (0.0%)	1 (25.0%)
benzene	11	5 (45.4%)	0 (0.0%)	1 (9.1%)
caprylene	19	5 (26.3%)	3 (15.8%)	5 (26.3%)
aminopyrazine carbaldehyde	2	2 (100.0%)	0 (0.0%)	2 (100.0%)
valienamine	7	1 (14.3%)	0 (0.0%)	1 (14.3%)

Table IX – Continued

Ligand Fragment	Number Examples	Accuracy Combined Features	Leave Fold Family Out	Remove Homologous Sequences
ethylsulfanyl isocyanatoethane	4	1 (25.0%)	0 (0.0%)	1 (25.0%)
difluorobenzyl alcohol	5	2 (40.0%)	1 (20.0%)	1 (20.0%)
picoline	6	2 (33.3%)	0 (0.0%)	2 (33.3%)
trihydroxy methyl				
aminohexanal	4	1 (25.0%)	0 (0.0%)	1 (25.0%)
beta arabino				
furanosylamine	79	36 (45.6%)	27 (34.2%)	37 (46.8%)
iodopyrazole	6	2 (33.3%)	0 (0.0%)	2 (33.3%)
isoequilenin	16	10 (62.5%)	2 (12.5%)	10 (62.5%)
aminomethyl methylpyrrolo				
pyrimidinone	6	1 (16.7%)	0 (0.0%)	0 (0.0%)
amino difluorohydroxy				
methylheptanal	7	2 (28.6%)	0 (0.0%)	0 (0.0%)
amino hydroxyindane	15	10 (66.7%)	0 (0.0%)	6 (40.0%)
fluoromethylbenzene	11	2 (18.2%)	0 (0.0%)	2 (18.2%)
aminomethyl pyrimidine	3	2 (66.7%)	0 (0.0%)	2 (66.7%)
methylpyridinone	3	1 (33.3%)	0 (0.0%)	1 (33.3%)
271	3	2 (66.7%)	0 (0.0%)	2 (66.7%)
ethanimidoyl piperidinol	5	3 (60.0%)	0 (0.0%)	2 (40.0%)
tertbutyl ethyl				
carbamate	8	1 (12.5%)	0 (0.0%)	1 (12.5%)
butane	9	2 (22.2%)	0 (0.0%)	2 (22.2%)
butyl alcohol	212	166 (78.3%)	158 (74.5%)	166 (78.3%)
cyclopentanamine	14	1 (7.1%)	1 (7.1%)	1 (7.1%)
methylpyrrolidinol	3	3 (100.0%)	3 (100.0%)	3 (100.0%)
hydroxyethylmethyl				
imidazole carboxamide	9	7 (77.8%)	0 (0.0%)	7 (77.8%)
methyltrifluoromethyl				
pyrimidinone	14	8 (57.1%)	3 (21.4%)	8 (57.1%)
amino methyl				
pyridinone	2	1 (50.0%)	2 (100.0%)	1 (50.0%)
ethanimidoylcyclo				
hexanamine	7	1 (14.3%)	1 (14.3%)	1 (14.3%)
ethyl methylthiazole	9	3 (33.3%)	0 (0.0%)	3 (33.3%)
aminoethyl benzene				
sulfonic acid	19	5 (26.3%)	0 (0.0%)	2 (10.5%)
diamino quinazolinone	13	2 (15.4%)	0 (0.0%)	2 (15.4%)
acetamido methyl				
boronic acid	6	4 (66.7%)	0 (0.0%)	4 (66.7%)
flavin	131	122 (93.1%)	114 (87.0%)	123 (93.9%)

Table IX – Continued

Ligand Fragment	Number Examples	Accuracy Combined Features	Leave Fold Family Out	Remove Homologous Sequences
aminomethyl benzimidazolyl	63	46 (73.2%)	6 (9.5%)	38 (60.3%)
methylidene azanium		17 (65.4%)	0 (0.0%)	17 (65.4%)
methionine		5 (55.6%)	0 (0.0%)	4 (44.4%)
thiamin		4 (57.1%)	3 (42.9%)	3 (42.9%)
cobinamide dihydrate	7			
Total	6830	5748 (84.2%)	5354 (78.4%)	5705 (83.5%)
p-value			< 0.0001	0.0002

Both Tables VIII and IX indicate that the interaction patterns between proteins and ligands is conserved more across fold family than across homologous sequences. Initial automated analyses of protein function were based on sequence patterns alone [116]. But, very soon these were complemented by secondary structure as well as fold family information since the geometric patterns in active sites were found to be essential to the understanding of protein function [111], [76]. In this study, we have also argued that over evolution, proteins use diverse sets of amino acids in the active sites to effect the same electrostatic patterns. The empirical results in Tables VIII and IX show that the classification accuracy decreases significantly when members of the same fold family are not considered during feature-matching and this decrease is not as significant when sequences with high homology are removed and this result agrees with the traditional knowledge within the field as well as our starting assumption.

These tests eliminating members of the same fold family and eliminating homologous sequences were performed to show that the classification methodology developed in this study is robust and that it is independent of fold family similarity as well as sequence homology. In a real-world application of this methodology to the functional analysis of a newly-solved protein structure, the training examples belonging to its fold family or those with high sequence homology will not be eliminated.

H. Large Active Site Analysis

Another way to characterize the accuracy of the classification algorithm is to analyze its performance on multi-fragment ligand active sites. In order to characterize the accuracy, it is first necessary to define the active site and generate the feature-based description of the active site. Active sites were defined around the entire ligand instead of the individual fragments by cutting out a 10Å pocket from the protein molecular surface centered at the centroid of the ligand. For each of the mesh points in this larger active site, a 5Å subpocket was then drawn centered at each of these mesh points and each of these subpockets was described using the position-dependent stereochemical features. The assumption in dividing the larger active site into subpockets is that the individual fragments of the ligand are centered on one of the mesh points in the active site and finding the centers of the fragments will allow for the identification of the identity of the multi-fragment ligand.

There are two metrics that will help better understand the accuracy of the feature-based methodology. The first is the relationship between fragment classification accuracy and the distance of the mesh point from the fragment centroid. The expectation is that as the mesh points are closer to a particular fragment centroid, the feature difference between the actual fragment pocket and the test subpocket will be small but as we move away from the fragment centroid, this feature difference will increase. The interaction patterns should be the clearest when the subpocket center is closest to fragment center since the fragment pockets in the database are all centered on the fragment centroid. As the distance from the centroid increases, the interaction patterns will include more noise and also information from regions of the active site not related to a particular fragment (possibly buffer regions or parts of active site interacting with other fragments etc) The feature difference is the *Euclidean* distance

between feature vectors (defined in Equation 5.7) and is directly correlated with the posterior probability obtained from the *KDE* classifier and therefore follows a similar pattern. Figure 21 shows the variation of feature difference with the distance from fragment centroid and shows that the feature difference does increase with distance from fragment centroid. At the same time, when a mesh point is within 1.5Å of the actual fragment centroid, the features of the subpocket show great similarity to the actual fragment pocket. This observation is encouraging since it ensures that a slight difference in subpocket centroid and fragment centroid is handled within the system and this proves that the feature-based descriptions are robust to small amounts of noise (due to interaction patterns from other regions).

The *Euclidean* distance between feature vectors X and Y is given by $d_{X,Y}$ is given as

$$d_{X,Y} = \sqrt{\sum_{i=1}^K (X_i - Y_i)^2} \quad (5.7)$$

where K is the length of the feature vectors.

Another metric to understand accuracy is a *site-wide* accuracy. If an active site houses a multi-fragment ligand the expectation is that the strongest probability peaks across the entire site will be related to one or more of the fragments of the ligand. For example, if an active site binds *PLP*, if the highest probability peaks (from *KDE* classification of mesh points) across all mesh points were sorted, the fragment classes *pyridoxal* and *phosphate* should have the highest peaks since the interaction patterns observed in the site should be most similar to those observed for the individual fragments of *PLP*. The following formula was used to determine the rank of each ligand fragment, f_i , in a ligand, L :

$$R(f_i) = R(P_j(f_i)) \quad \forall j = 1 : N \quad (5.8)$$

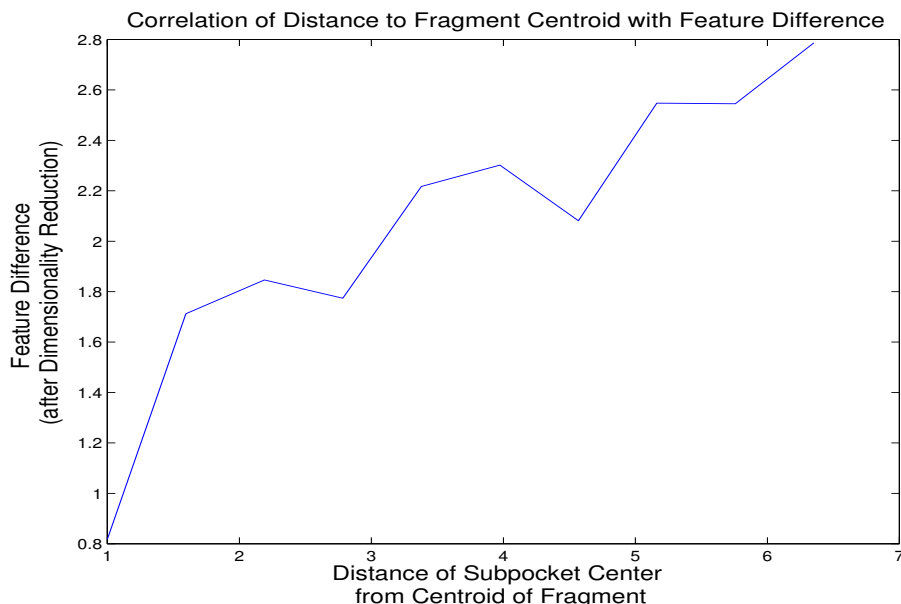


Fig. 21. The variation of feature difference with distance from actual fragment center for all the fragment classes in this study

where $R(f_i)$ is the rank of the i^{th} fragment of ligand L and $P_j(f_i)$ is the likelihood of fragment f_i at mesh point j obtained from the classification algorithm (Equation 2.32) and $R(P_j(f_i))$ is the rank of this probability value over all the N mesh points in the active site. Site-wide accuracy for each of the fragment classes in a subset of the multi-fragment ligands was computed and the results of this experiment are summarized in Table X.

Table X.: Site-wide accuracy

Ligand Name	Number of Examples	Avg. Rank of Fragments Across Examples	Average number of mesh points
2DT	13	thymine:7.0, ribose:2.3, phosphate:1.1	212
ADP	84	adenine:2.6, ribose:2.3, diphosphate:4.8	202
AMP	29	adenine:2.3, ribose:2.6, phosphate:1.1	210
ANP	18	adenine:2.3, ribose:2.7, triphosphate:15.7	195
ATP	46	adenine:2.4, ribose:2.6, triphosphate:13.8	222

Table X – Continued

Ligand Name	Number of Examples	Avg. Rank of Fragments Across Examples	Average number of mesh points
FMN	60	flavin:13, butyl-alcohol:9.0, phosphate:0.7	199
PLP	171	pyridoxal:4.9, phosphate:1.1	168
PMP	12	pyridoxal:5.3, phosphate:0.9	173
SAH	27	adenine:2.9, ribose:2.2, methionine:42.1	201
SAM	15	adenine:2.7, ribose:2.2, methionine:26.4	208
TDP	14	thiamin:134.6, diphosphate:4.9	222
TMP	24	thymine:6.0, ribose:2.6, phosphate:1.0	205
UDP	16	thymine:5.9, ribose:2.4, diphosphate:5.7	225

Table X shows that in almost all cases the correct fragments (those belonging to the multi-fragment ligand binding the active site) are within the top 10 fragments with the highest probabilities. Considering that the database contains examples from 441 different fragment classes, these results are highly encouraging. Fragment *thiamin* from *TDP*) is ranked the lowest (average rank of 134), but considering there are only 9 examples of *thiamin* in this database with 7068 fragments, the lower rank is not surprising. It is encouraging to note that despite the very low number of examples, the active sites binding *TDP* still showed any peaks at all for *thiamin* (average probability of 0.015). For example, in 17/29 *AMP* binding sites, there was at least one *thiamin* peak with an average probability of 0.008. In fact, on average, *thiamin* peaks were found in greater than 50% of binding sites that do not bind *thiamin* and the average probability of *thiamin* in these sites was found to be 0.01. The difference in the average probability of *thiamin* in sites that actually do bind *thiamin* (0.015) and those that do not (0.01) is obviously not significant enough to believe in either the presence or absence of *thiamin*. It is in such cases that combining the geometric information regarding the fragment placement in multi-fragment ligands will allow to discriminate between spurious peaks for a ligand fragment and those that characterize

fragment binding.

The data regarding the above two metrics combined with the fragment accuracy calculations demonstrate that the position-dependent stereochemical features do indeed capture the interaction patterns of the ligand fragments with the protein and that the dimensionality reduction techniques as well as the classification scheme are all able to distinguish clearly between the 441 fragment classes analyzed in this study.

I. Combination of Fragments Using Markov Random Field

The feature-based classifier provides the probability of individual fragments within a given test active site. But, only a subset of these also satisfy the geometric constraints on the placement of these individual fragments. The geometric constraints are obtained as defined in Section C in Chapter III. The mean and standard deviations between the fragments for a subset of the ligands in this study is listed in Table XI.

Table XI.: Statistical models for the distances between various fragments in the larger ligands

Ligand Name (L)	Fragment Name, a	Fragment Name, b	$\mu_{L,a,b}$	$\sigma_{L,a,b}$
2DT	Ribose	Phosphate	4.085	0.444
2DT	Thymine	Phosphate	6.076	1.150
2DT	Thymine	Ribose	3.510	0.903
ADP	Adenine	Diphosphate	7.227	1.258
ADP	Adenine	Ribose	4.106	0.173
ADP	Ribose	Diphosphate	5.202	0.498
AMP	Adenine	Phosphate	6.572	1.166
AMP	Adenine	Ribose	4.097	0.177
AMP	Ribose	Phosphate	4.446	0.374
ANP	Adenine	Ribose	4.162	0.209
ANP	Adenine	Triphosphate	8.055	1.328
ANP	Ribose	Triphosphate	5.938	0.634
ATP	Adenine	Ribose	4.114	0.141
ATP	Adenine	Triphosphate	7.929	1.204
ATP	Ribose	Triphosphate	5.940	0.655
FMN	Flavin	FAD-Carb	5.051	0.170
FMN	Flavin	Phosphate	8.173	0.966
FMN	Phosphate	FAD-Carb	4.168	0.376
PLP	Pyridoxal	Phosphate	4.600	0.643
PMP	Pyridoxal	Phosphate	4.376	0.555
SAH	Adenine	Homocysteine	8.236	1.226
SAH	Adenine	Ribose	4.106	0.170
SAH	Ribose	Homocysteine	6.000	0.593
SAM	Adenine	Homocysteine	8.258	1.334
SAM	Adenine	Ribose	4.101	0.178
SAM	Ribose	Homocysteine	5.920	0.591
TDP	Thiamin	Diphosphate	7.263	1.059

Table XI – Continued

Ligand Name (L)	Fragment Name, a	Fragment Name, b	$\mu_{L,a,b}$	$\sigma_{L,a,b}$
TMP	Ribose	Phosphate	4.290	0.325
TMP	Thymine	Phosphate	5.526	0.881
TMP	Thymine	Ribose	3.323	0.253
UDP	Ribose	Diphosphate	5.182	0.498
UDP	Thymine	Diphosphate	6.726	1.143
UDP	Thymine	Ribose	3.609	0.427

The simulated annealing procedure described in Section D in Chapter III is used to find fragment combinations based on the geometric constraints as well as the individual fragment class posterior probabilities from the *KDE* classifier. The joint probability of each fragment combination is evaluated and the fragment combinations are ranked by this probability. Multi-fragment ligands are clustered such that all ligands that are completely contained in another multi-fragment ligand are grouped together. For *e.g.*, ligand *G6P* is composed of fragments *ribose* and *phosphate* and since both these fragments are contained in ligand *AMP*, ligands *G6P* and *AMP* are grouped together. The multi-fragment ligands with the top 100 probability values are examined and a count of the number of times each multi-fragment ligand cluster is represented within the top 100 is determined. The sorted list of these counts is then analyzed to determine the final labeling for the active site. Table XII lists the ranking within the top 100 for a subset of the multi-fragment ligands in this study. This table shows that in 8 out of 13 cases the true ligand was within the top 10 ligands when ranked by the count metric.

Table XII.: Results of *mrf* analysis

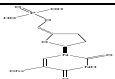
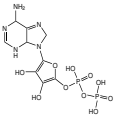
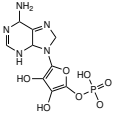
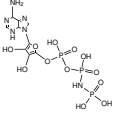
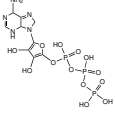
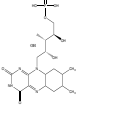
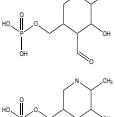
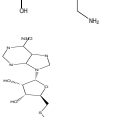
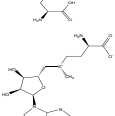
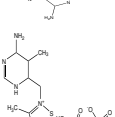
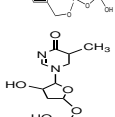
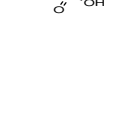
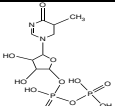
Ligand	Ligand Name	Number of Examples	Number of Fragments	Number Correctly Labeled	Rank
	2DT	13	3	13	9.15
	ADP	84	3	84	9.86
	AMP	29	3	29	8.55
	ANP	18	3	18	7.4
	ATP	46	3	46	7.26
	FMN	60	3	59	86.69
	PLP	171	2	171	4.0
	PMP	12	2	12	4.0
	SAH	27	3	27	34.22
	SAM	15	3	15	33.8
	TDP	14	2	14	53.35
	TMP	24	3	24	8.83

Table XII – Continued

Ligand	Ligand Name	Number of Examples	Number of Fragments	Number Correctly Labeled	Rank
	UDP	16	3	16	21.63

An examination of the results for the multi-fragment ligand *FMN* in Table X and Table XII showed that while the active sites for *FMN* showed significant peaks for the individual fragments (*flavin*, *butyl-alcohol* and *phosphate*), the average rank for *FMN* after the *Markov* combination is very low suggesting that these peaks did not fit the geometric constraints for *FMN*. Similarly, the results for *TDP* indicate that while the active site did not show significant peaks for *thiamin*, the existing peaks fit the geometric patterns for *TDP* thereby improving the overall rank of *TDP* in the final list of fragment combinations.

J. Test Cases

In the previous sections, the accuracy of the algorithm on the database containing 7070 patches from 2310 unique proteins complexed with 1160 unique ligands was analyzed. In order to ensure that the accuracy of the proposed methodology on this database is not due to any bias in training, three test complexes were selected and the accuracy of the algorithms were tested on these complexes. All three test cases have functional annotations in the *PDB*, but the complex structure was not solved for one example (*1qde*) and there are very few structural/sequence homologs for the other two test cases. In all three cases, there exist biochemical studies based on sequence analysis of the proteins that suggest their cognate ligands and mode of

activity. Figure 22 shows the steps involved in the analysis of a test protein based on the methodology developed in this study.

1. *DEAD* Box Protein: 1qde

The first test case is *1qde*, a structure of the ATPase domain of the translation initiation factor 4A (eIF4A) [10]. *eIF4A* is a prototype of the *DEAD* box protein family and proteins within this family are involved in cellular processes like cellular splicing, ribosome biogenesis and RNA degradation. *eIF4A* melts the local secondary structure of RNA and makes it more susceptible to nucleases in the presence of an energy source, *ATP*. While many proteins have been characterized as *DEAD* box proteins based on sequence motifs (Asp-Glu-Ala-Asp), enzymatic activity has been confirmed in only a subset of these proteins [77]. Since the interaction with *ATP* is essential for enzymatic activity, the active site will be analyzed for interaction patterns related to ATP binding. This protein belongs to the *SCOP* fold family *P-loop containing nucleoside triphosphate hydrolases* which is a common *ATP* binding fold but *SCOP* also notes that the *P-loop* in *1qde* has a non-canonical conformation, thus making it challenging to characterize the function of this protein.

Active Site Definition: Benz *et.al.* [10] identified amino acids that interact with ligands *ADP* or *AMP* based on structural comparison with other *DEAD* box proteins. Based on this analysis, they identified that Phe41 would potentially make hydrophobic van der Waals interactions with the *adenine* moiety and Glu43 and Gln48 would form hydrogen bonds with the *adenine* atoms. The *phosphate* moiety would interact with residues Gly68-Thr72. The authors mentioned that they did not observe any interaction between the *ribose* moiety and the protein. In a similar analysis, the protein *2vso*, a complex of a RNA helicase from yeast with *AMP* was identified as the closest structural and sequence homolog using *DALI* [54] and *BLAST* respectively.

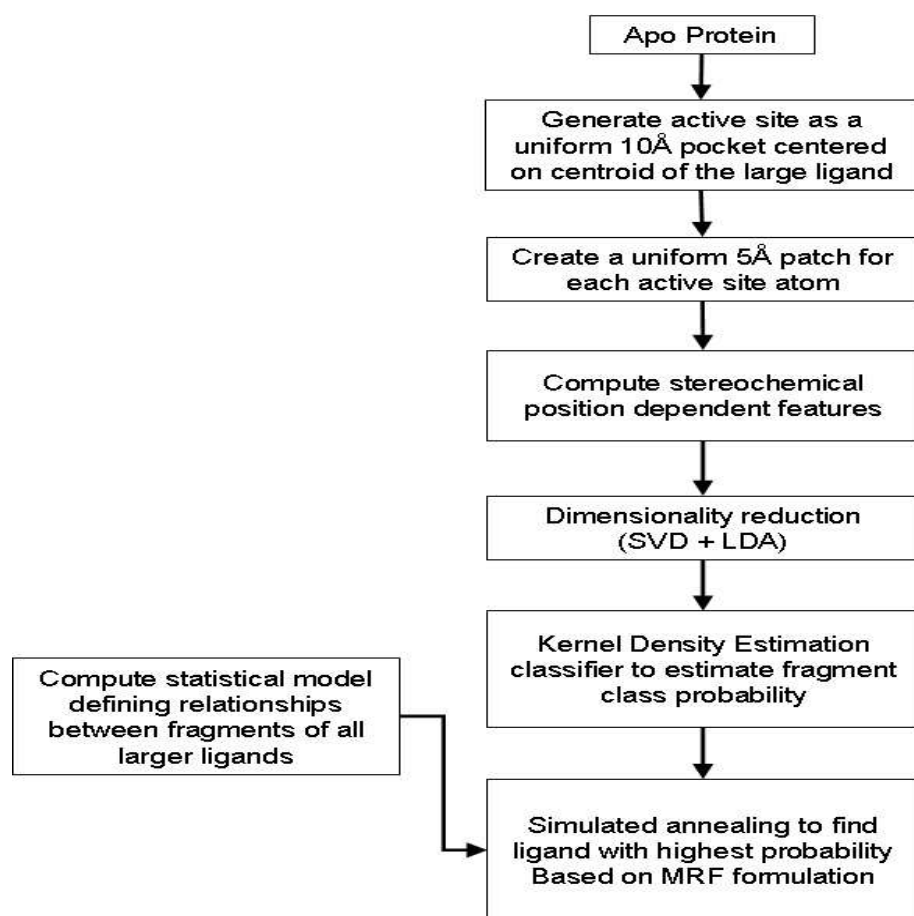


Fig. 22. Flowchart showing the various steps involved in the analysis of a test protein

The superposition of *1qde* and *2vso* is consistent with the above mentioned protein-ligand interactions and Figure 23 shows these interactions with *1qde* superposed onto *2vso*. Based on this information, the active site was defined as a 10Å patch of the

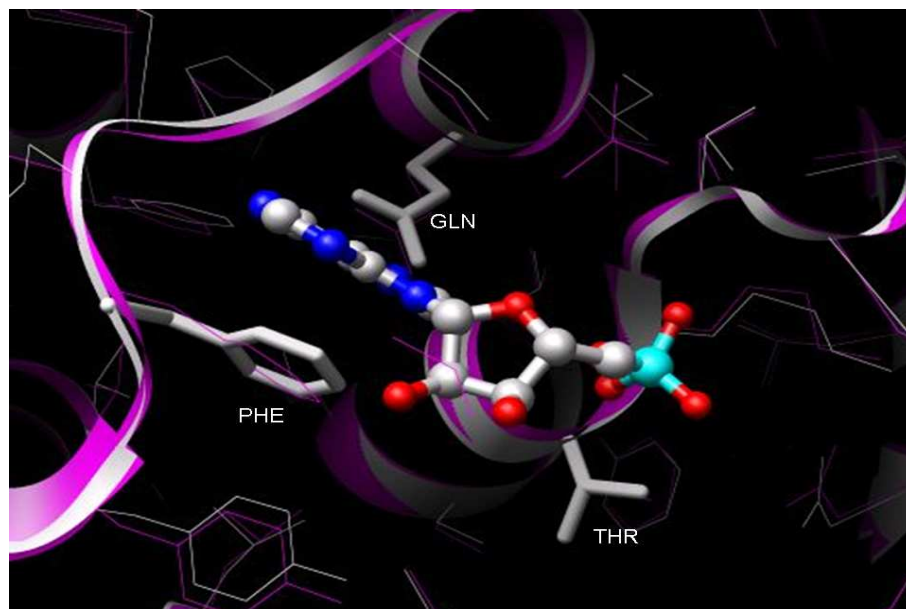


Fig. 23. The active site residues identified by superposing the structure of *1qde* (shown in white) with the structure of *2vso* complexed to the ligand *AMP*

protein molecular surface centered at the center of the above mentioned residues. The molecular surface was generated based on the procedure outlined in Chapter II. Interestingly, the dimer structure of *2vso* shows an interaction between the second protein domain and the ribose of *AMP* (shown in Figure 24) suggesting a reason why there are no interactions between (a single chain of) *1qde* and ribose. Figure 25 shows the active site derived for this protein.

Analysis of a test active site: 5Å patches centered at each of the 356 mesh points within the active site were generated. Position-dependent stereochemical features were determined for each 5Å patch and the dimensionality of these features was

reduced using the combined *SVD+LDA* technique described previously in Chapter II and each mesh point was annotated with the 441 class labels from the database, each associated with a probability score from the *KDE* described in Chapter II. Figures 26(a), 26(b) and 26(c) show the mesh points labeled as *adenine*, *ribose* and *phosphate* respectively with high posterior probability (from *KDE*). These figures showed that while a few mesh points were incorrectly labeled, there exists a clear concentration of these peaks near the correct fragment. For *e.g.*, in Figure 26(a) there are a few mesh points closer to the *ribose* and *phosphate* fragments but the majority of mesh points labeled *adenine* are clustered near the placement of the *adenine* fragment.

These posterior probability values were combined with the geometric constraints using the *mrf* formulation described in Chapter III and various possible combinations of labellings across the active site were evaluated. The labeling combinations were sampled based on simulated annealing techniques again described in Chapter III. The labellings with the highest probability for each multi-fragment ligand in the database were listed. Based on the multi-fragment ligand clusters determined above in Section I, the multi-fragment ligand clusters most represented within the top 100 labellings from the results of the simulated annealing procedure were determined. The 10 most represented clusters and the counts for these clusters as well as the probability associated with each multi-fragment ligand for *1qde* are tabulated in Table XIII.

Table XIII.: Results of *mrf* analysis for *1qde*

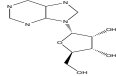
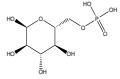
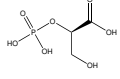
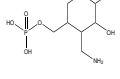
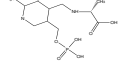
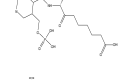
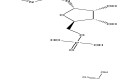
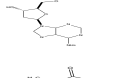
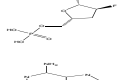
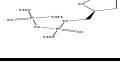
Ligand	Ligand Name	Cluster Count	Probability of fragment combination from <i>mrf</i>
	ADN	35	0.0003

Table XIII – Continued

Ligand	Ligand Name	Cluster Count	Probability of fragment combination from <i>mrf</i>
	G6P	26	0.0005
	2PG	26	0.0008
	PMP	15	2.05e-05
	PP3	9	3.76e-08
	KAM	9	7.21e-08
	IMP	9	4.73e-07
	DTP	9	3.33e-08
	FDM	8	1.19e-07
	ADP	8	2.87e-07

5/10 of the ligands in this list are connected to *AMP*, *ADP* and *ATP* (*ADN* (containing *adenine* and *ribose*) and *G6P* (containing *ribose* and *phosphate*) are both completely contained within the ligand *AMP*, ligand *IMP* has exactly the same fragments as ligand *AMP*, ligand *DTP* has exactly the same fragments as ligand *ATP*). *DEAD* box proteins are known to bind all three of these ligands and therefore the final labeling obtained from the methodology outlined in this study has been successfully able to identify the function (determine that it binds *ATP*) of the protein

1qde.

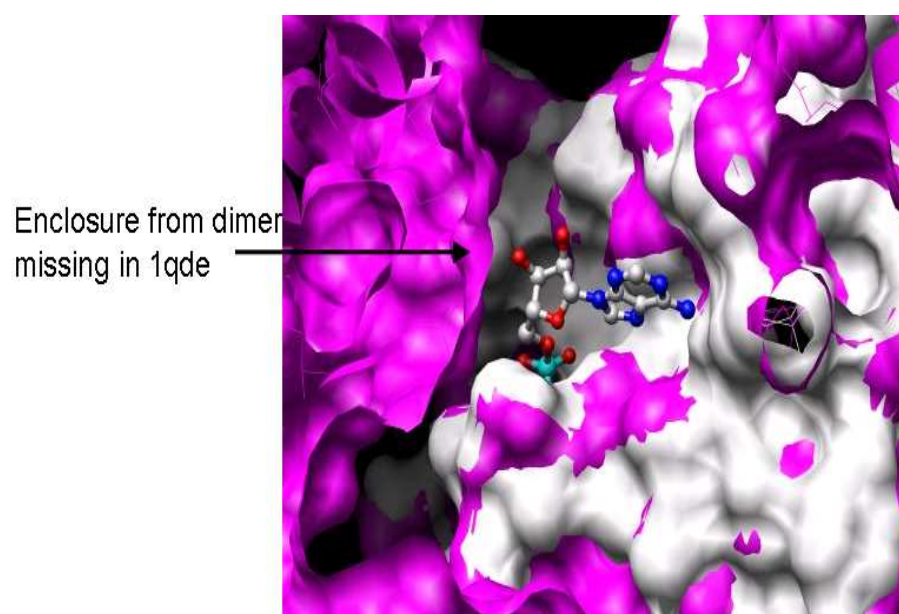


Fig. 24. The dimer structure of *2vso* shows the dimer plays a role in defining the interaction of *ribose* with the protein

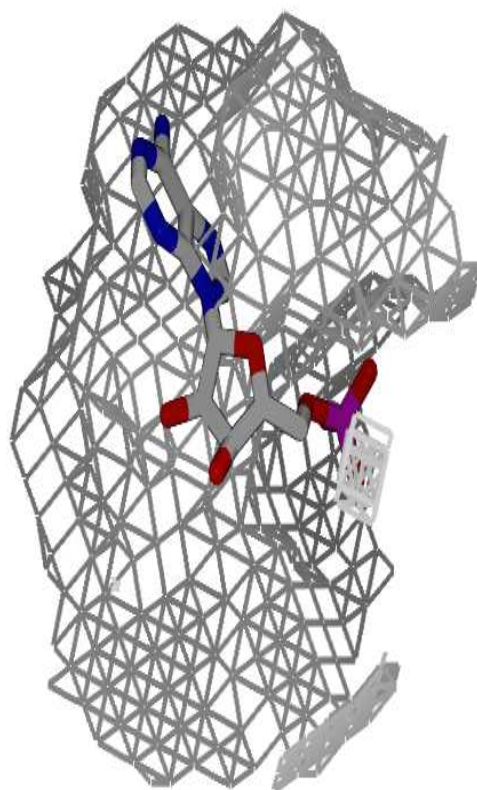
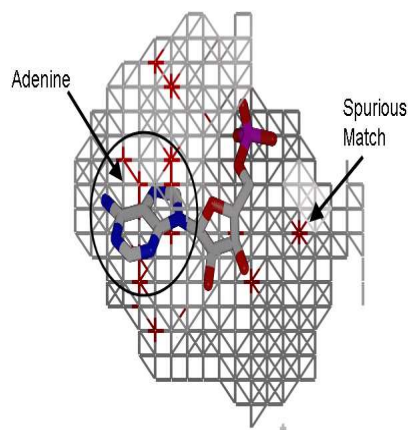
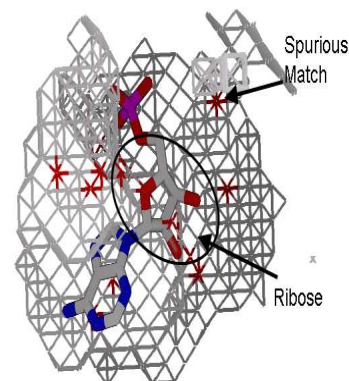


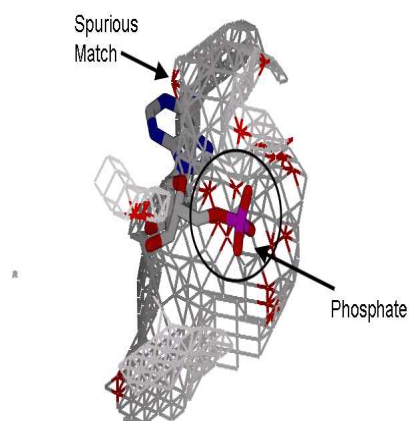
Fig. 25. The active site for *1qde* with the *AMP* structure from *2vso*



(a) The mesh points with high probability of being centers of the *adenine* fragment highlighted in red



(b) The mesh points with high probability of being centers of the *ribose* fragment highlighted in red



(c) The mesh points with high probability of being centers of the *phosphate* fragment highlighted in red

Fig. 26. Figures 26(a), 26(b) and 26(c) show the peaks for the various fragments of ligand *AMP* based on the classification algorithm based on the stereochemical features computed at each of the mesh points. While the majority of the peaks are near the actual fragment centers, there are a few spurious matches spread throughout the active site

2. *PriA* Protein: 2d7h

The second test case is *2d7h*, the structure of the *PriA* protein from *E. coli*. The 3D structure of this protein is complexed with ligand *deoxycytidine monophosphate* (dCMP). It has been characterized as a *primosomal* protein based on biochemical analyses. A *BLAST* search of this protein sequence yielded three matches, one of which is the apo structure of this protein *2d7e* with an E-value of 2e-56 and the other two matches were not significant with E-values of 3.6 and neither of these proteins is present in our database. The fold of this protein has not been characterized by *SCOP* and a search for structural homologs using *SSM* of the entire *SCOP* database and using *DALI* to search the entire *PDB* yielded no structurally similar proteins. If the complex structure of this protein was unavailable, the above-mentioned characteristics would make it difficult to analyze the function of this protein based on sequence or fold similarity analyses.

Active Site Definition: To date no structural studies detailing the active site of this protein have been published. In this case, since the complex structure of the protein was available, the residues interacting with the ligand *dCMP* were used to identify the active site. The residues lining the active site are Phe16 (chain A&B), Thr15 (chain B), Glu41 (chain B) and Ile43 (chain B). The active site residues of *2d7h* with *dCMP* bound to it is shown in Figure 27 and the molecular surface around the protein outlining the active site surface is shown in Figure 28. Based on this information, the active site was defined as a 10Å patch of the protein molecular surface centered at the center of the above mentioned residues. The molecular surface was generated based on the procedure outlined in Chapter II.

Analysis of a Test Active Site: 5Å patches centered at each of the 487 mesh points within the active site were generated. Position-dependent stereochemical fea-

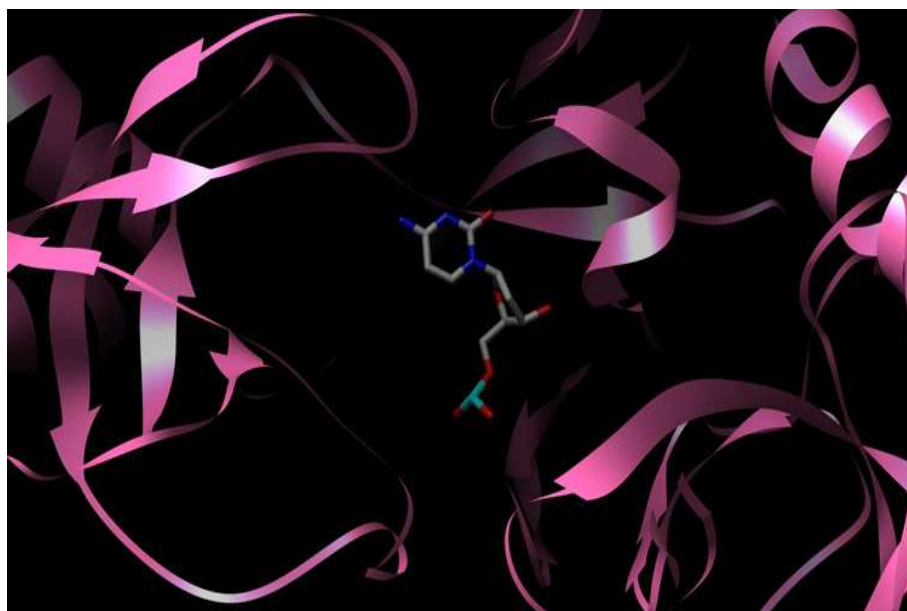


Fig. 27. The active site residues interacting with the ligand *dCMP*

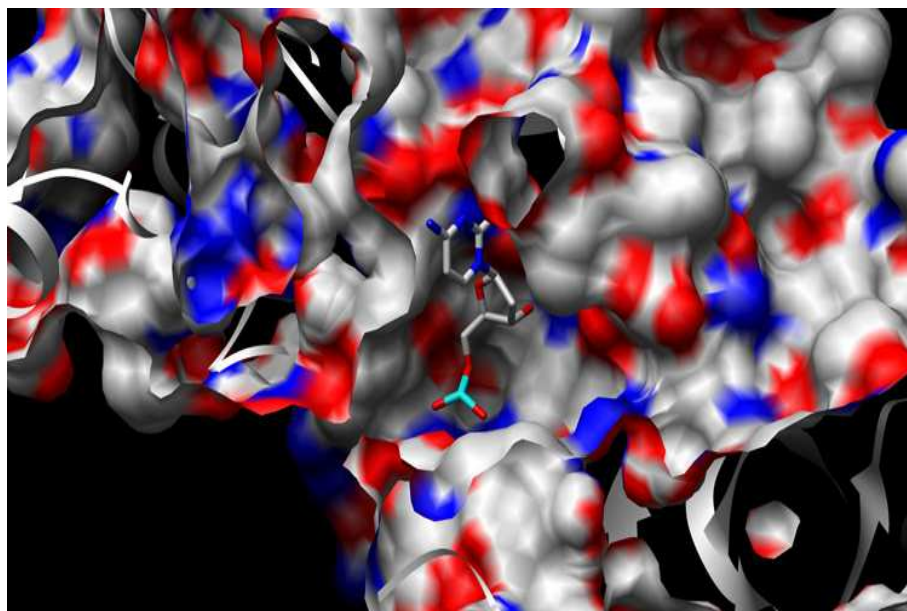
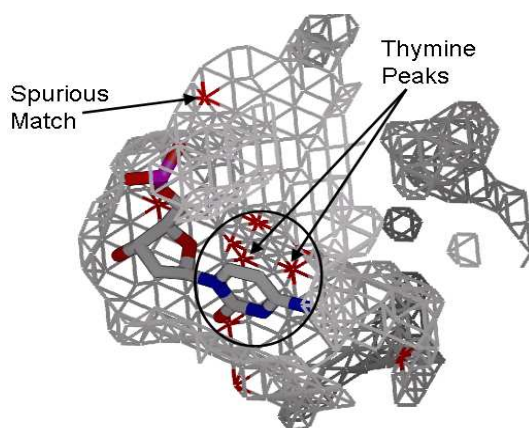


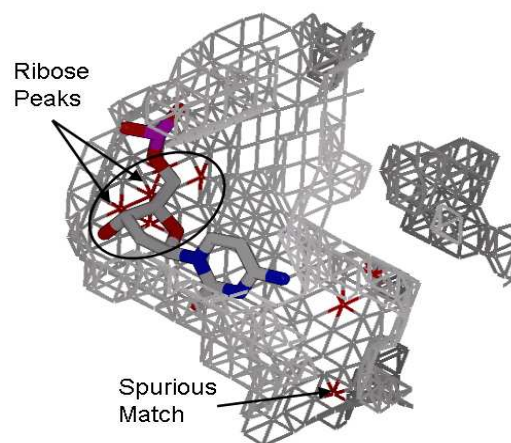
Fig. 28. Molecular surface of *2d7h* showing *dCMP* in the active site

tures were determined for each 5Å patch and the dimensionality of these features was reduced using the combined *SVD+LDA* technique described previously in Chapter II and each mesh point was annotated with the 441 class labels from the database, each associated with a probability score from the *KDE* described in Chapter II. Figures 29(a), 29(b) and 29(c) show the mesh points labeled as *thymine*, *ribose* and *phosphate* respectively with high posterior probability (from *KDE*). These figures showed that while a few mesh points were incorrectly labeled, there exists a clear concentration of these peaks near the correct fragment. For *e.g.*, in Figure 29(a) there are a few mesh points closer to the *ribose* and *phosphate* fragments but the majority of mesh points labeled *thymine* are clustered near the placement of the *thymine* fragment.

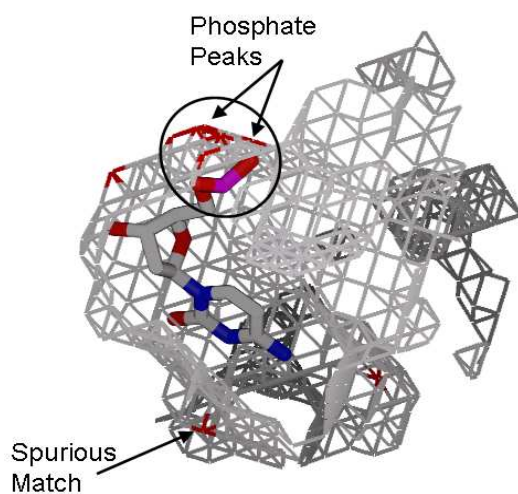
These posterior probability values were combined with the geometric constraints using the *mrf* formulation described in Chapter III and various possible combinations of labellings across the active site were evaluated. The labeling combinations were sampled based on simulated annealing techniques again described in Chapter III. The labellings with the highest probability for each multi-fragment ligand in the database were listed. Based on the multi-fragment ligand clusters determined above in section I, the multi-fragment ligand clusters most represented within the top 100 labellings from the results of the simulated annealing procedure were determined. Table XIV lists the ranks of all the ligands in the clusters related to *dCMP* and the corresponding probabilities.



(a) The mesh points with high probability of being centers of the *thymine* fragment highlighted in red



(b) The mesh points with high probability of being centers of the *ribose* fragment highlighted in red



(c) The mesh points with high probability of being centers of the *phosphate* fragment highlighted in red

Fig. 29. The distribution of the peaks for each of the fragment classes within ligand *dCMP* through the entire active site. Despite some spurious peaks the majority of the peaks are clustered around the centers of *dCMP* fragments

Table XIV.: Results of *mrf* analysis for *2d7h*

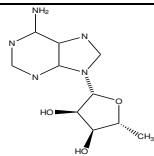
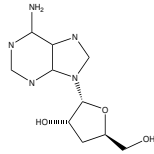
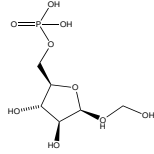
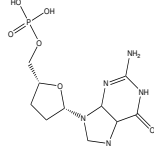
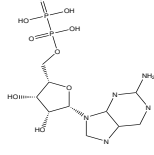
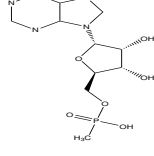
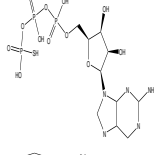
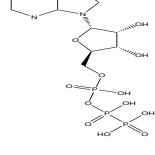
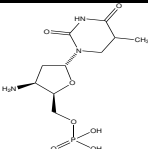
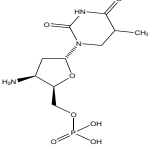
Ligand	Ligand Name	Rank	Cluster Count	Probability of fragment combination from <i>mrf</i>
	5AD	1	48	0.0003
	3AD	2	48	0.022
	F6P	3	36	0.0005
	DGP	6	14	6.04e-07
	AP2	7	14	0.0003
	ABM	8	14	0.04
	ATG	9	11	0.0003
	ACP	10	11	0.0002

Table XIV – Continued

Ligand	Ligand Name	Rank	Cluster Count	Probability of fragment combination from <i>mrf</i>
	NYM	11	10	1.71e-05
	C31	12	10	0.002

Ligands *5AD*, *3AD*, and *F6P* are all completely contained within the ligand *dCMP*. Ligands *DGP*, *AP2* and *ABM* are all the same as the ligand *AMP* and since *thymine* and *adenine* are related bases, finding these ligands within the top 10 clusters is highly encouraging. Ligands *ATG* and *ACP* are the same as the ligand *ATP*, a ligand very similar to *AMP* and once again observing these ligands in the top 10 is encouraging. Finally, *NYM* and *C31* are the same as *dCMP* and are ranked within the top 13 multi-fragment ligand clusters (out of 764 possible ligands). These results show that there is a very strong signal in this active site for the ligand *dCMP* even in the absence of any fold similarity or sequence homology of this protein with all the example proteins in the database.

3. Hypothetical Protein PA1024: 2gjl

The third and final test case is *2gjl*, the first ever crystal structure of *2-nitropropane dioxygenase*. This 3D structure is for the protein from *Pseudomonas aeruginosa* complexed with the ligand *flavin mononucleotide* (FMN). It is a 328 residue protein with 23% sequence identity with the *nitropropane dioxygenase* from *N. crassa* [47]. The authors of this structure did not find any significant structural similarity between this protein and all other proteins in the PDB using *DALI*. The *SCOP* fold of this protein is also uncharacterized.

Active Site Definition: Ha *et.al.* [47] identified that amino acids Gly22, Gln24, Thr75, Lys124, Asp145, Ala150, Ser178, Gly180, Gly201, and Thr202 interact with the ligand *FMN* in the active site (shown in Figure 30). The phosphate moiety of FMN, buried completely inside the pocket, is not solvent-accessible, whereas the edge of the isoalloxazine ring is partially accessible from the protein surface (shown in Figure 31). They also found that Gly180, Gly201, and Thr202 constituted the standard phosphate binding motif also utilized by other members of the *FMN-dependent oxidoreductases* and *phosphate-binding* enzymes. Based on this information, the active site was defined as a 10Å patch of the protein molecular surface centered at the center of the above mentioned residues. The molecular surface was generated based on the procedure outlined in Chapter II.

Analysis of a Test Active Site: 5Å patches centered at each of the 409 mesh points within the active site were generated. Position-dependent stereochemical features were determined for each 5Å patch and the dimensionality of these features was reduced using the combined *SVD+LDA* technique described previously in Chapter II and each mesh point was annotated with the 441 class labels from the database, each associated with a probability score from the *KDE* described in Chapter II. Figures

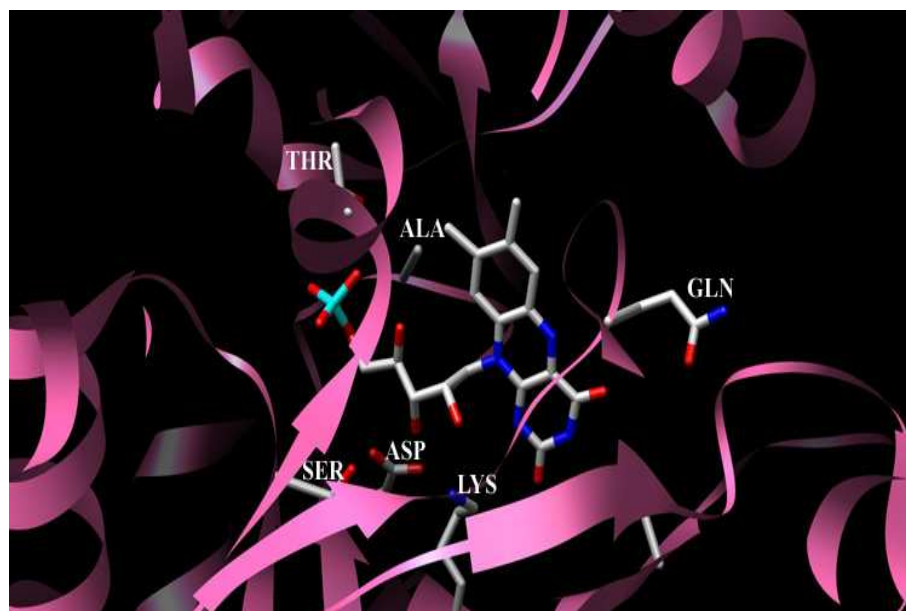


Fig. 30. The active site residues interacting with the ligand *FMN* as identified by [47]

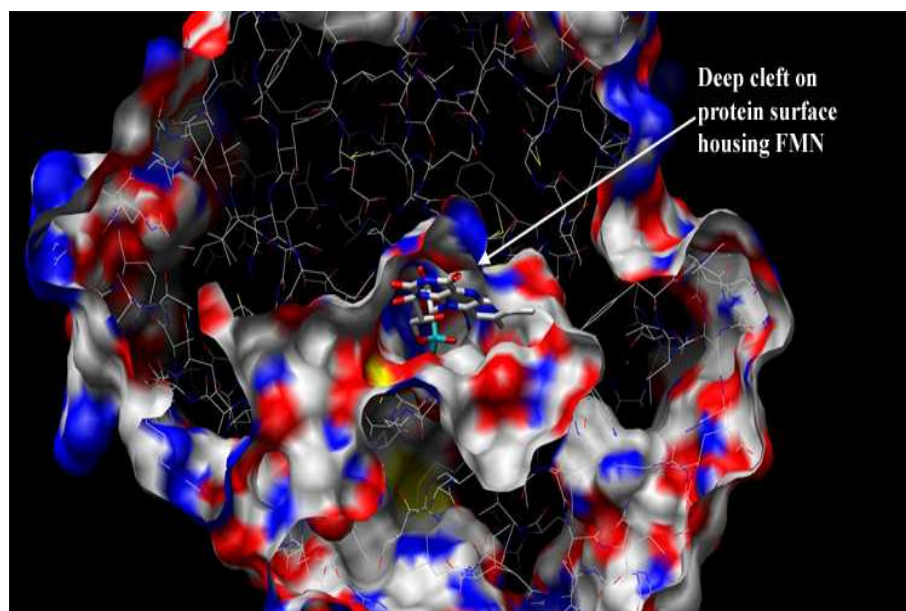


Fig. 31. Molecular surface of *2gjl* showing the deep cleft housing *FMN*

32(a), 32(b) and 32(c) show the mesh points labeled as *flavin*, *butyl-alcohol* and *phosphate* respectively with high posterior probability (from *KDE*). These figures showed that while a few mesh points were incorrectly labeled, there exists a clear concentration of these peaks near the correct fragment. For *e.g.*, in Figure 32(a) there are a few mesh points closer to the *butyl-alcohol* and *phosphate* fragments but the majority of mesh points labeled *flavin* are clustered near the placement of the *flavin* fragment.

These posterior probability values were combined with the geometric constraints using the *mrf* formulation described in Chapter III and various possible combinations of labellings across the active site were evaluated. The labeling combinations were sampled based on simulated annealing techniques again described in Chapter III. The labellings with the highest probability for each multi-fragment ligand in the database were listed. Based on the multi-fragment ligand clusters determined above in section I, the multi-fragment ligand clusters most represented within the top 100 labellings from the results of the simulated annealing procedure were determined. No multi-fragment ligands related to *FMN* were in the top 10 of these clusters. Table XV lists the ranks of all the ligands in the clusters related to *FMN* and the corresponding probabilities.

Table XV.: Results of *mrf* analysis for *2gjl*

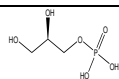
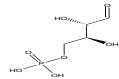
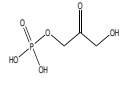
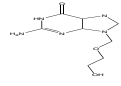
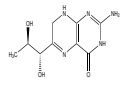
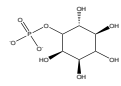
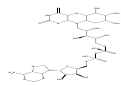
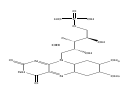
Ligand	Ligand Name	Rank	Cluster Count	Probability of fragment combination from <i>mrf</i>
	G3P	22	5	1.95e-05
	E4P	23	5	1.94e-05

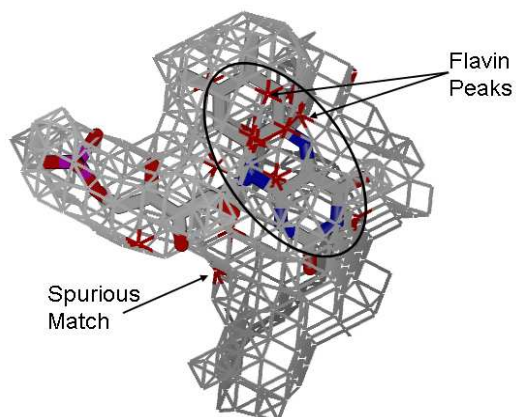
Table XV – Continued

Ligand	Ligand Name	Rank	Cluster Count	Probability of fragment combination from <i>mrf</i>
	13P	24	5	1.59e-05
	AC2	25	5	1.45e-05
	HBI	26	5	1.11e-05
	IPD	27	5	8.83e-06
	6FA	47	4	1.17e-24
	FMN	67	1	1.74e-09

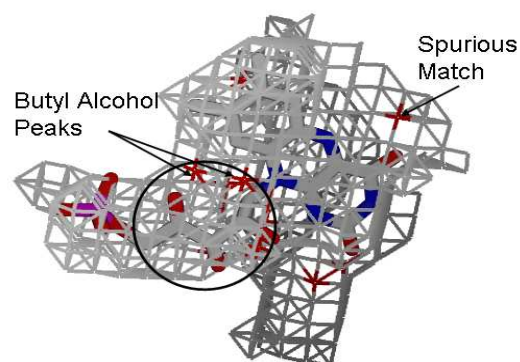
The results in the table show that there are multiple ligands (6) within the top 50 clusters which are completely contained within *FMN*. The first 6 ligands listed in Table XV all contain *butyl-alcohol* and *phosphate*, 2 fragment classes in *FMN*. Additionally, the ligand *FMN* is completely contained within ligand *6FA*. While, the correct ligand is not within the top 10 clusters, the correct ligand and similar multi-fragment ligands are listed within the top 50 (out of 764 possible multi-fragment ligands).

K. Effect of Protonation States

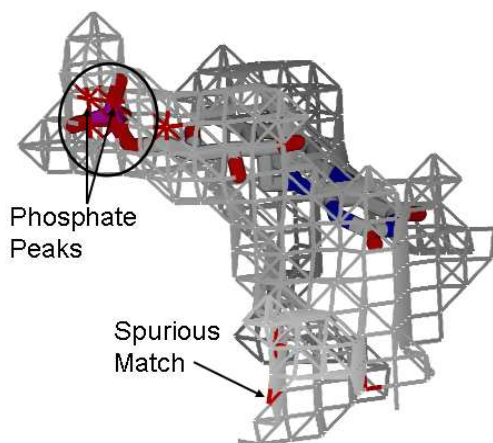
As mentioned in Chapter II, the protonation states were not considered in this study. To study the effects of this choice, an experiment was conducted with one protein



(a) The mesh points with high probability of being centers of the *flavin* fragment highlighted in red



(b) The mesh points with high probability of being centers of the *butyl-alcohol* fragment highlighted in red



(c) The mesh points with high probability of being centers of the *phosphate* fragment highlighted in red

Fig. 32. The distribution of the peaks for each of the fragment classes within ligand *FMN* through the entire active site. Despite some spurious peaks the majority of the peaks are clustered around the centers of *FMN* fragments

(Dictyostelium discoideum myosin motor domain; 1VOM bound to *ATP*) that has protonated histidines in the active site. One set of electrostatic features was calculated without considering the protonation states and another set using the protonation states identified by *PROPKA*. These two sets of features are then used in classification as well as to find the multi-fragment ligand bound to the protein. The aim was to ascertain if using the protonation states would have an effect on classification and if so, identify the extent of the effect.

A biochemical study of 1VOM by *Lawson et.al* [71] identified histidines 12, 297, 408 and 550 as protonated residues. *PROPKA* also identified these residues as protonated and the pK_a values of 7.09, 6.63, 7.21 and 6.43 respectively. Based on these pK_a values, these histidines were considered as positively charged and the charge associated by *AMBER* with protonated histidines were used in the determination of the electrostatic features.

Dimensionality reduction was performed on both sets of electrostatic feature vectors using the previously described methodologies. The reduced dimension vectors were then used to classify the three pockets associated with the fragments of the ligand *ADP* (*adenine*, *ribose* and *phosphate*). Table XVI shows the differences in classification accuracy using the protonated version of the feature versus the deprotonated version for all three pockets.

Table XVI.: Differences in classification accuracy using protonated versus deprotonated versions of the electrostatic feature vector

True Class	Probability using	
	Deprotonated	Protonated
Electrostatic Features		
Adenine	0.025	0.0003
Ribose	0.225	0.00005
Phosphate	0.125	0.001

This comparison shows clearly that including the protonation states greatly changes the active site chemistry, a fact clearly borne out by the large difference in the classification accuracies of the three ligand fragments. This decrease in classification accuracy is directly due to the differences in the protonated and deprotonated versions of the electrostatic features. In this particular case, the head-to-head comparison is unfair to the protonated version of the feature vector since all the other examples of *ADP* binding sites in the database are deprotonated. This shows that any future inclusions of protonation states need to be applied consistently to the entire database.

L. Application to Drug Discovery

The study of structure-function relationship in proteins is primarily aimed at aiding the development of successful drugs. Traditional drug development approaches are based on high-throughput screens of large chemical compound libraries, but these methodologies have not been as successful as expected. Recent studies ([66], [131]) have shown that the leads returned by these screens were on average 4-5 heavy atoms larger than the corresponding drugs showing that the compound libraries currently being used for screens contain really large molecules. Starting with a large initial scaffold also makes it harder to optimize the leads. Fragment-based drug screens are an alternative paradigm in drug discovery aimed at the identification of small, low molecular weight drug fragments that interact with the active site. Using the fragment based-approach has the potential to keep the overall complexity of each drug candidate low allowing for better lead-optimization techniques.

The fragment-based methodology developed in this dissertation for functional

analysis has many parallels to the fragment-screen approach. Both methodologies use fragments to better capture the local interactions between the protein and ligand/chemical compound. Additionally, previous fragment screening approaches like *TETHERING* [33] and *GROMOL* [14] have also used the steric fitness of fragments to help grow the initial fragments to get the final drug in a manner very similar to the steric constraint based *MRF* combination developed in this study. Therefore, it is possible to extend our methodology to aid in the drug discovery process.

In order to extend our methodology we need to develop a database of interactions between proteins and small molecule fragments similar to the current protein-ligand fragment interaction database. The steric constraints between the small-molecule fragments can be captured using a model very similar to the one we currently use and the combination techniques using the *MRF* formulation will remain the same. Unfortunately, given the vastness of chemical space, it is highly unlikely that the structure of a protein bound to each one of the small molecules has been determined. Considering that a typical small-compound library has thousands of compounds, it would be very difficult to capture the diversity of these interactions using a small set of protein-inhibitor complexes. On the other hand, it might be possible to cluster the small molecule space into similar chemotypes and ensure that there is at least one example of each cluster in the database. The success of our methodology in modeling the protein-ligand interactions increases the possibility of success for this approach of drug design and development. This is certainly an approach that merits future analysis and experimentation.

M. Comparison to Previous Methods of Active Site Analysis

In this section, we will compare the feature-based classification methodologies developed in this study to the current available feature-based methods. This analysis is difficult because not all the studies use the same datasets or evaluation methods. Nonetheless, we will run our method against methods developed by Gutteridge *et al* and Denessiouk *et al*, using the data described in their paper, and determine the accuracy of classification by our definition, in order to assess their relative performance.

A previous study by Gutteridge *et al* [46] analyzed active sites using a feature-based methodology. The authors used features based on residue identity and placements to characterize active sites. The active site was characterized as a 12Å surface patch around the coordinates of the ligand (the authors do not specify how the active site would be characterized in the absence of the ligand). The specific features used to characterize the active sites detailed in [113] were:

Residue hydrophobicity: A hydrophobicity scale [35] was used to determine the hydrophobicity of each of the residues within the active site.

Residue solvent accessibility: *NACCESS* [56] was used to find the solvent accessibility of each of the atoms within the protein. The solvent accessibility of the atoms within a residue was summed to find the residue solvent accessibility.

Residue secondary structure: *DSSP* [59] is used to determine the secondary structural elements present within the protein and the placement of residues within each secondary structural element can be obtained.

Residue b-factor: The *b-factor* values associated with each of the atoms in the residue are obtained from the *pdb* file and summed to find the *b-factor* value of the entire residue. This value provides information regarding the flexibility observed for

each of the residues and this information is especially useful for the residues within the active site.

Residue identity: The similarity between residues is based on the *BLOSUM* matrices.

A subset of our database consisting of 454 proteins bound to a diverse set of ligands was chosen. The active site was defined as all the residues with any atom within 12\AA of any of the ligand atoms. Each of the features listed above was computed for each of these active site residues and feature vectors were created for each of these active site patches. Dimensionality reduction techniques were not applied to the feature vectors since the feature vector size was already small. A *nearest neighbor* algorithm was used to find the closest feature match for each of the active sites and 10-fold cross-validation was used in the analysis yielding a classification accuracy of 61.2%. As before, the proteins were additionally analyzed by removing all examples from the same fold family (51% classification accuracy) as well as by removing all examples with homologous sequences (45.8% classification accuracy). The significant reduction in classification accuracy after removing homologous sequences confirms that the features developed in this study rely predominantly on the residue identity similarities between active sites that bind the same ligand.

On the same dataset, the position-dependent stereochemical features were able to correctly identify the active sites binding similar ligands with a classification accuracy of 82.3% and once again removing examples from the same fold family made more of a difference to this analysis (79.3%) while removing homologous sequences did not make as much of a difference (80.9%). This analysis shows that the position-dependent stereochemical features are better at capturing active site patterns between diverse proteins binding the same ligand as opposed to the residue identity features previously

developed.

In a 2001 study of proteins of ancient origin bound to ATP, CoA, FAD, NAD and NAP by Denessiouk *et al* ([28], [27]), the adenine binding site was characterized by sequence motifs corresponding to polar interactions and hydrophobic interactions, both of which were not specific to sequence or fold families. In their study, they were unable to characterize all adenine binding sites using their motif. The motif developed in this study was applied to a set of 500 adenine active sites and the motif was only able to characterize 65% of them accurately. The localized stereochemical features were also applied to the same set of adenine active sites and were able to accurately classify the adenine active sites with greater success ($> 90\%$) despite the diversity in the protein fold and sequences between these example active sites.

Both the above examples show that the methodology of localized and position-dependent stereochemical features developed in this study go beyond traditional sequence identity and feature description methodologies and are able to successfully identify the similarities in active site stereochemistries despite the diversity observed within active sites binding the same ligand.

Figure 33 shows the comparison of classification accuracies using the methodologies developed in this study, the feature-based methodology (Gutteridge *et al*).

N. Conclusions and Future Work

In this chapter, the results of the active site analysis methodologies developed in this dissertation were presented. The approach presented in this study was based on a feature-based description of active site interaction patterns between proteins and ligands. The discussion in Section F showed that stepping away from the traditional global features and utilizing more granular stereochemical features allows to capture

the interaction patterns in greater detail and with greater fidelity, ultimately yielding greater classification accuracy. The results of analyzing the localized versus the position-dependent features showed that the position-dependent features performed much better than the localized features (by correctly categorizing more fragment classes and an increased accuracy of 8%).

The feature-based classifier was also robust when proteins belonging to the same fold family as well as those with sequence homology ($> 35\%$) were ignored. Ignoring the members of the fold family had a greater effect on classification accuracy (5%) as opposed to ignoring members with high sequence homology (1-2%). This result confirms expert knowledge in this field regarding the importance of conserving 3D structure over conserving sequence and further increases the belief in the accuracy of the methodology used in this study.

When analyzing a test active site, it is necessary that the classification algorithm be robust to slight errors in the definitions of the fragment pockets since very often the test active site is defined approximately based on the position of known conserved residues or possible interactions. The analysis in Section F showed that while the feature difference increased as the distance between the actual fragment center and the test pocket centroid increased, there was nevertheless a buffer of 1.5\AA where the feature difference was not as high to preclude the possibility of correctly classifying the test pocket. This ensures that slight variations in active site definition will not have an adverse effect on the results using the feature-based analysis presented in this study.

Analyzing the test active site based on the fragment class probabilities obtained from the *KDE* classifier confirmed that most often the individual fragments of the multi-fragment ligand had the highest peaks through the active site proving that the feature-based classifier was indeed capturing the interaction patterns through the

active site between protein and ligand fragments.

Finally, the combination of the ligand fragments into a final ligand prediction using *mrf* combination shows the power of using contextual information (in this case in the form of geometric constraints on fragment placements within a ligand) to combine the probability of observing individual fragments into probabilities of different combinations of those individual fragments. In each of the test cases, the true ligand was within the top 10 ligands as ranked by the final combination probability.

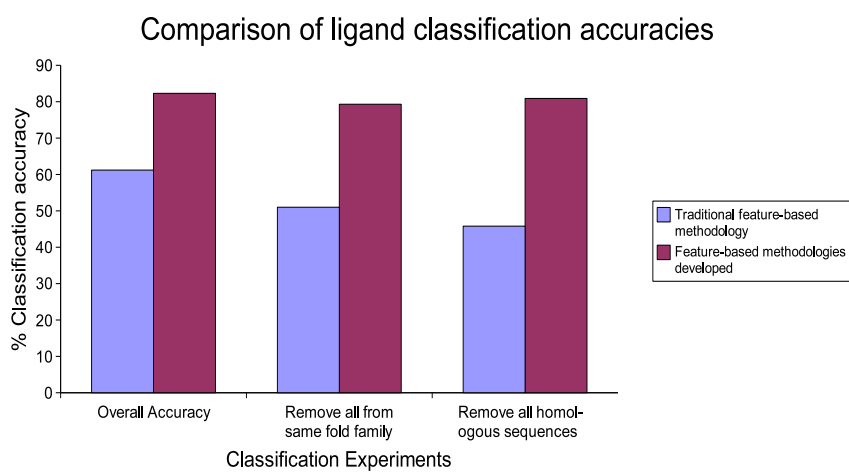


Fig. 33. Comparison of classification accuracies using the methodologies developed in this study, the feature-based methodology (Gutteridge *et al*)

CHAPTER VI

SPECIFICITY NORMALIZATION FOR IDENTIFYING SELECTIVE
INHIBITORS IN VIRTUAL SCREENING

Virtual screens offer yet another avenue to functionally characterize an apo-protein and they are being increasingly used to complement high throughput screening (HTS) of large chemical libraries in the search for hits and leads for many newly-solved protein 3D structures [19], [43], [38]. Unfortunately, the results of virtual screens are often populated by many compounds that have no activity against the target receptor but are ranked higher than known inhibitors. In this chapter, I will introduce a novel ranking scheme, *Rscore*, intended to improve the enrichment and recall of known inhibitors and thereby increase the probability of finding other novel inhibitors using computational methods. *Rscore* minimizes the number of false positives by taking into account the interaction patterns of known inhibitors across a variety of decoy active sites. It then normalizes the *DOCK* score of a compound based on this pattern using a linear programming formulation. *Rscore* was tested on two different target receptors and showed an increase in recall of most of the known inhibitors by greater than 20%. I will also present the results of experimental validation of *Rscore* on the *Malate Synthase* (MS) receptor. Laboratory experiments based on *Rscore* ranking led to the testing of 16 compounds (ranked within top 50 of 250,000 compounds); of which 4 were identified as micromolar inhibitors for MS. I will also present an analysis of the compounds that are consistently ranked at the top of multiple virtual screen runs in order to identify and characterize *virtually promiscuous* compounds.

A. Previous Approaches

Despite the successes of virtual screening approaches, many deficiencies still exist with this methodology. The majority of docking algorithms are unable to handle the flexibility in receptors due to induced fit (though some programs can account for limited receptor flexibility [1], [103]). More importantly, the scoring functions used in various docking algorithms can only approximate the protein-ligand/small-molecule interaction energy due to the various approximations and trade-offs involved in their formulations. Since these functions are key to ranking the docked ligand/small molecule poses in large-scale virtual screening runs, very often, the final interaction score for known inhibitors does not compare favorably to the scores of other *drug-like* compounds that do not show any inhibition experimentally. The success of a virtual screen is largely defined by how often the methodology yields novel heretofore unexplored chemical scaffolds for a given receptor [115]. Most docking algorithms do not claim to rank these interesting chemotypes higher than uninteresting/non-specifically interacting compounds. This definition of virtual screen success underscores the need to better rank the results of a virtual screen. Accurate ranking is especially needed when analyzing large screening libraries, since human analysis of each small-molecule interaction with the receptor becomes less feasible.

Scoring functions estimate interaction energies in many different ways, ranging from empirical force fields (with typical electrostatic and van der Waals terms) [78], to statistical force fields (*e.g.* PMF [88]), and some try to account for the effects of solvation and ligand conformation [15]. *Stahl et. al* [121] empirically compared 4 different scoring functions (*FlexX*, *PLP*, *DrugScore* and *PMF*) across 7 different receptor sites and found that each scoring function, because of its formulation as well as the parameters used, performed better on certain classes of small-molecules

(lipophilic, polar etc.). None of them was able to perform well on a large and diverse database, thereby significantly reducing the usefulness of these scoring functions in large-scale virtual screens. Consensus scoring schemes have often been suggested as a way to combine individual scoring functions [128]. Consensus score ranks compounds by using multiple scoring functions and chooses the worst rank based on these multiple scoring functions. Implementations of consensus scoring often drop one or two of the worst scores in order to give compounds a fair chance (account for scoring function biases). The consensus score seeks to select molecules that are consistently ranked higher with each of the individual scoring functions. Unfortunately, this scoring scheme is typically found to be only as successful as the best scoring function used [22], [92], [121].

Stahl et. al [121] also defined *ScreenScore* as a linear combination of the 4 scoring functions mentioned above and found that while it did not perform as well as the *PLP* and *FlexX* scoring functions on 2 of their 7 receptor sites, they observed an improved performance against the other sites. Since the new score was a linear combination of the previous scores, it was able to evaluate a diverse range of compounds with higher accuracy, thereby increasing the diversity within the virtual screen results.

Despite these incremental improvements in the scoring function formulations, the ranking of known inhibitors in the results of a virtual screen often remains low due to the presence of a large number of false positives (small molecules with large negative interaction energies but no observable biochemical inhibition) in this list. The different scoring functions defined till date have been focussed on evaluating the interaction between a given receptor and a small-molecule. Typical scoring functions do not take into account the *specificity* of interaction with the receptor, relative to other receptors. It is quite possible that some compounds have high interaction energies with multiple active sites due to the bias inherent in the scoring functions.

For example, *DOCK* tends to consistently rank large and charged compounds much higher than other compounds (data shown below). Since the aim of virtual screening is to identify small molecules that have specific and significant interactions with the receptor site, it is essential to include this specificity analysis when ranking the results of a virtual screen. In large libraries with 10^6 - 10^7 compounds, if the known inhibitors are not ranked within approximately the top 1%, there may be thousands of false positives with apparently good docking scores that must be assayed before finding those with true inhibitory activity.

In our recent work [97], we presented a novel approach that increased the recall and enrichment rate of virtual screens by improving the ranking of known inhibitors versus non-inhibitors. We defined a ranking function *Rscore* that takes into account the specificity of the small-molecule's interaction with the protein by calibrating the *Dock* score to the target receptor against scores from docking to functionally different active sites (*decoy sites*). We employed a linear programming formulation and determined a set of weights for the interaction of the molecule to the decoy sites in order to optimize the *Rscore* value for known inhibitors versus those for non-inhibitors. In our experiments, we used DOCK6.1 [86] as the docking algorithm and the *DOCK* score (or Grid energy) as the initial scoring function. The small-molecules from the ChemBridge drug-like library were used as the database in these experiments.

In this chapter, we extend this work and experimentally validate the *Rscore* ranking on *Mycobacterium tuberculosis* malate synthase and show that *Rscore* is indeed able to identify novel inhibitors as well as novel scaffolds of interaction between enzyme and small molecules. Our experiments on *Rscore* revealed some small molecules that were consistently ranked at the top of multiple virtual screen runs. In this chapter, we will identify and characterize these *virtually promiscuous* small molecules.

B. Methods

In this section, we will present the mathematical basis of *Rscore*. *Rscore* seeks to evaluate the specificity of interaction between a target receptor and small molecules by comparing the *DOCK* score to the target receptor against the scores to the decoy sites. The basic assumption behind the approach is that a good inhibitor should have a large negative score against the target receptor and have lower magnitude interaction energies against the decoy sites. *Rscore* re-ranks the results of a virtual screen by incorporating more information about interactions with decoy sites so as to increase the percentage of known inhibitors at the top of the ranked list.

Rscore takes into account three factors in its re-ranking scheme: (a) the relative rank of a compound against the target receptor in comparison to its average rank across the decoy sites; where rank is computed in ascending order of *DOCK* scores, (b) the number of times *DOCK* fails (for example, when a compound does not fit into the receptor site), and finally (c) the number of times a compound docks with a positive score (due to insufficient conformational sampling). Each of these three conditions (ranks based on docking with a negative score, docking with a positive score or not docking at all) reflects the “dockability” of the small-molecule in different ways, and *Rscore* seeks to combine this information.

Let $P_1...P_n$ define the n decoy active sites and P_0 define the target receptor and $r_0..r_n$ are the ranks based on docking scores to each of the receptors (as defined above). Then *Rscore* can be written as

$$Rscore = w_1\delta + w_2\pi + w_3\phi \tag{6.1}$$

where δ is the difference between rank based on *DOCK* score to target receptor vs. mean rank over decoys. δ is calculated based on μ_d the average rank over decoy sites

$$\delta = r_0 - \mu_d \quad (6.2)$$

$$\mu_d = \frac{1}{n} \sum r_{d_i} \quad (6.3)$$

$$\pi = \text{number of receptors with positive scores} \quad (6.4)$$

and

$$\phi = \text{number of docking failures} \quad (6.5)$$

We seek weights w_1 , w_2 and w_3 so as to minimize the *Rscore* value for inhibitors as compared to the *Rscore* value for non-inhibitors. The choice of the weights is crucial to the correct ranking of known inhibitors and non-inhibitors. Non-inhibitors can be sampled randomly from the small-molecule library, assuming most of the compounds from the library do not have any inhibition activity. In this study, we use linear programming to find a set of weights that maximizes the number of times the known inhibitors are ranked higher than non-inhibitors.

1. Linear Programming Formulation

In the linear programming formulation, constraints are defined and the most stringent constraints can be written as

$$\begin{aligned} Rscore_i - Rscore_j &\geq 0 \quad \forall i \in \text{non-inhibitors}, \\ &\quad \forall j \in \text{inhibitors} \end{aligned} \quad (6.6)$$

where $Rscore_i$ and $Rscore_j$ are the values of *Rscore* (as defined by Equation 6.1) for a non-inhibitor and an inhibitor respectively. Substituting Equation 6.1 into Equation 6.6 and rearranging the terms we get

$$w_1(\delta_{non} - \delta_{inh}) + w_2(\pi_{non} - \pi_{inh}) + w_3(\phi_{non} - \phi_{inh}) \geq 0 \quad (6.7)$$

Multiple such constraints can be defined by repeatedly randomly choosing a non-inhibitor and a known inhibitor, computing their values of δ , π and ϕ and finally, formulating a constraint as in Equation 6.7. Since, there are likely to be some inhibitors and/or some non-inhibitors that do not meet the above defined constraints, slack variables are introduced into each constraint and the weights w_1 , w_2 and w_3 are chosen such that the sum of these slack variables is minimized (reducing the number of times a non-inhibitor is ranked higher than an inhibitor). This less stringent constraint is written as

$$Rscore_i - Rscore_j + s_k \geq 0 \quad (6.8)$$

where s_k defines the slack variables introduced into each constraint and k runs over the number of constraints created. The linear program formulation is written as

$$Minimize : \sum_{k=1}^C s_k \quad (6.9)$$

$$s.t. \quad w_1 + w_2 + w_3 = 1 \quad (6.10)$$

$$s.t. \quad Rscore_i - Rscore_j + s_k \geq 0 \quad k = 1 : C \quad (6.11)$$

where C is the total number of constraints.

2. Enzyme Assay for Malate Synthase

A coupled assay that monitored the release of CoA from the MS-mediated reaction was utilized for MS inhibition. The assay was carried out using 100 μ L overall reaction volumes with MS at 92.5 nM being reacted in 20 mM Tris pH 7.5 and 5 mM $MgCl_2$. All inhibitors (in 100% DMSO) were added such that the final reaction mixture contained

1% DMSO; 1 μ L of stock inhibitor added to the reaction mixture. Inhibitors were incubated with MS in tris buffer with MgCl_2 for 10 min at room temperature before adding 0.625 mM acetyl CoA, the first reaction was initiated by the addition of 1.25 mM of glyoxylate, the second substrate.

The coupled assay measures the increase in absorbance at 412 nm due to the formation of 5,5'-dithiobis-(2-nitrobenzoic acid) (DTNB)-CoA adduct. DTNB is injected with glyoxylate at the reaction starting point. A BMG LABTECH POLARstar OPTIMA plate reader in absorbance mode was used to continuously monitor the reaction for 2 min per well of a Corning 96-well plastic plate. Reaction with a 1% DMSO solution instead of inhibitor was taken as the uninhibited control. The percent inhibition was calculated by comparing the slope/min values (representing the enzyme velocity) of an inhibitor trial to the uninhibited control.

C. Results

The performance of *Rscore* was tested on three different enzymes: COX-2, dihydrofolate reductase (DHFR) and *Mycobacterium tuberculosis* (*Mtb*) malate synthase. Ideally, the decoy sites should be chosen such that they have low sequence homology and structural similarity with the receptor as well as among themselves so as to capture a diverse set of receptor environments. The 9 decoy active sites used in this study are *Mtb* alanine racemase, 1XFC (Alr [72]), *Mtb* type II dehydroquinase, 1H0R (AroD [109]), diaminopelargonic acid synthase, 3BV0 (BioA [30]), *Mtb* 1-Deoxy-D-xylulose 5-phosphate reductoisomerase, 2JCZ (DXR [49]), *Mtb* long fatty acid chain enoyl-ACP reductase, 1ZID (InhA [110]), *Mtb* malate synthase, 1N8W (MS [117]), *Mtb* pantothenate synthetase, 2A7X (PanC [129]), *Plasmodium falciparum* enoyl-acyl-carrier-protein reductase, 1NHG (PfENR [100]) and *Mtb* phosphoglycerate dehydro-

genase, 1YGY (PGDH [31]). Each of these active sites was defined based on the coordinates of the bound ligands as well as published active site definitions. The receptors were all prepared by adding hydrogens and applying AMBER charges [25] using Sybyl [124].

The 250,000 drug-like small molecules from the ChemBridge library (<http://www.chembridge.com>) were docked into each of these active sites using Dock6.1. These small molecules were prepared using Openeye software [94] by adding hydrogens and applying Gasteiger charges. It is assumed that none of these small-molecules show any inhibition against the target receptors and therefore these molecules are used as examples of non-inhibitors (negative examples of inhibitors) in future linear programming formulations.

In each of these experiments, 100 constraints were created by randomly picking a non-inhibitor and an inhibitor and adding a constraint as defined in Equation 6.8. The known inhibitors were used in training (to optimize the weights using the linear programming formulation) and also used in testing (to evaluate whether the use of *Rscore* improves the ranking of known inhibitors). Therefore, care was taken to ensure separation between training and test cases by using a *leave-one out* method. N-1 known inhibitors are used for creating the constraints and determining the optimal weights and the remaining inhibitor is used as test case. For each set of 100 constraints, the values for w_1 , w_2 and w_3 were obtained using *GLPSOL* available as part of the GNU Linear Programming Kit (<http://www.gnu.org/software/glpk>). This was repeated 300 times and the final set of weights was defined as the average of the weights obtained in each linear programming iteration.

The following subsections list a detailed analysis of the results of *Rscore* rankings for each of the three enzymes.

1. *Rscore* for COX-2

The COX-2 active site has been extensively studied and various NSAIDs (non-steroidal anti-inflammatory drugs) have been designed to interact with this receptor site [23], [62], [68], [87], [95], [98], [106], [107], [125]. We chose the specific 3D coordinates from 6COX (complexed with SC558) [68] to define the receptor site of interest. While multiple crystal structures exist for COX-II, only small differences in the active site conformations may be noted between them and therefore most of the inhibitors should dock to the chosen crystal structure (the conformations of Arg120 and Leu384 are the most varied, but these changes do not affect most inhibitors [34]).

Seventeen of the known inhibitors (arachidonic acid, Celebrex, Diclofenac, Etodolac, Etoricoxib, Flurbiprofen, Ibuprofen, Indomethacin, Ketoprofen, Lumaricoxib, Meloxicam, Naproxen, Piroxicam, Resveratrol, SC558, Valdecoxib and Vioxx) are listed in Figure 34. These known inhibitors form the set of positive examples used in this study.

Fourteen of the 17 known inhibitors docked successfully with negative *DOCK* score to the 6COX receptor site. Two of the known inhibitors (Valdecoxib and Vioxx) docked with positive scores and Indomethacin did not dock at all. The inhibitor (substrate) arachidonate had the highest (most negative) *DOCK* score (-60.69) and the inhibitor Etoricoxib has the lowest *DOCK* score (-22.77). Table XVII lists the *DOCK* score of the 17 known inhibitors against the 9 decoy active sites as well as the COX-2 site. This table shows that SC558 and Celebrex dock with a positive score in a majority of the decoy active sites (6/9 and 7/9 respectively) and do not dock against the remaining decoy sites. All the other inhibitors dock with a negative score (albeit lower *DOCK* score) with majority of the decoy sites. The weights obtained using the linear programming formulation for *COX-2* inhibitors were $w_1 = 0.98$, w_2

$= 0.0$ and $w_3 = 0.02$.

Table XVII.: *DOCK* scores of known COX-II inhibitors across various receptors

Inhibitor	Active Site									
	ALR	AroD	BioA	DXR	InhA	MS	PanC	PfENR	PGDH	COX-2
Arachidonate	599	-51.49	-64.98	-68.69	-38.42	-63.43	-57.95		-46.65	-60.69
Celebrex	2868		128	3851	10249	91.66		18.75	11068	-44.08
Diclofenac		720	-41.3		126	-35.74	-37.28	-41.10	10.79	-34.55
Etodolac	606	-21.92		2.69	-25.36	-37.46	-43.46	-42.30		-34.80
Etoricoxib		232	-31.68	238	-23.28	-32.39	3.73		61.17	-22.77
Flurbiprofen	87	-40.92	-34.46	-41.84	-19.13	-30.40	-40.79	-42.15	-11.78	-34.80
Ibuprofen		-43.9	-44.17	-41.66	6.67	-33.34	-41.07	-39.83	-15.35	-39.31
Indomethacin	1090	56.46	144	-40.68	-22.17	-35.26	-29.81	-53.53	2209	
Ketoprofen	19	-34.96	-40.3	-34.92	-26.27	-38.01	-41.03	-43.71	-13.01	-34.31
Lumaricoxib		197		-30.15	130	-20.69	-31.33	-36.63	1150	-20.36
Meloxicam	41	-20.77	-41.38	16.75	-38.41	-43.05			-7.63	-35.47
Naproxen	41	-39.17	-38.8	-42.34	-21.29	-32.88	-43.90	-39.84	-17.50	-43.72
Piroxicam	662	-29.92	-41.26	4.11	-38.43	-40.60	-39.75		19.40	-32.21
Resveratrol	-10.49	-32.15		-36.01	-46.40	-40.69	-34.52	-36.76	-20.61	-35.40
SC558	4468		100	3583	3662	61.34			249	-38.26
Valdecocixib	872	-22.31	-19.94	92.99	-30.49	34.95	8.05	-41.45	11.26	439
Vioxx	672	-16.03	-37.76	57.17	-43.54	-22.03	-33.60	-42.21	4637	71.92

The values of δ for each known inhibitor against the 9 decoy sites, the number of sites that have positive *DOCK* scores and the number of sites that the inhibitor fails to dock against are listed in Table XVIII. The value of *Rscore* is listed for each known inhibitor. The table lists the ranking of the inhibitor according to the original *DOCK* score and the ranking according to *Rscore*. It also lists the consensus score computed by finding the second worst rank based on *DOCK* score and CScore (Sybyl implementation that computes *DSCORE*, *PMFSCORE*, *GSCORE* and *CHEMSCORE*). This table shows that the ranking of most of the known inhibitors using *Rscore* greatly increases the enrichment rate; 7/14 rank within the top 10% and all 14 within the top 15%. Several increase in ranks by greater than 20%; *e.g.* Lumaricoxib increases in rank from 45% (*DOCK*) to the top 13% (*Rscore*). *Rscore* performs much better than the ranking using consensus score. Figure 35 compares the enrichment curves based on *DOCK* score, consensus score and *Rscore*.

Table XVIII.: *Rscore* calculation and its comparison to *DOCK* score. Ranks (shown in parantheses) are given as a percentage relative to the ChemBridge library containing 250,000 compounds. μ is the mean average rank over the decoy sites, δ is the difference between the rank against the target receptor and μ , π is the number of receptors with positive scores and ϕ is the number of decoy receptors with docking failures

Inhibitor	<i>DOCK</i> Score	μ	δ	π	ϕ	<i>Rscore</i>	<i>DOCK</i> Rank	Consensus Score Rank	<i>Rscore</i> Rank
Arachidonate	-60.69	0.18	-0.18	1	1	-0.15	95 (0%)	34197 (14%)	33605 (13%)
Celebrex	-44.08	0.83	-0.76	7	2	-0.71	13553 (1%)	96617 (37%)	3 (0%)
Diclofenac	-34.55	0.57	-0.30	3	2	-0.25	61191 (24%)	111463 (45%)	13808 (6%)
Etodolac	-34.8	0.56	-0.29	2	2	-0.25	59900 (24%)	52461 (21%)	14792 (6%)
Etoricoxib	-22.77	0.70	-0.21	4	2	-0.16	108064 (43%)	6907 (3%)	31761 (13%)
Flurbiprofen	-34.8	0.45	-0.18	1	0	-0.17	59913 (24%)	117320 (47%)	30160 (12%)
Ibuprofen	-39.31	0.40	-0.25	2	0	-0.24	34550 (14%)	146086 (58%)	14923 (6%)
Ketoprofen	-34.31	0.42	-0.14	1	0	-0.14	62412 (25%)	119320 (48%)	36576 (15%)
Lumiricoxib	-20.36	0.73	-0.20	3	2	-0.16	113706 (45%)	76150 (30%)	33080 (13%)
Meloxicam	-35.47	0.52	-0.27	2	2	-0.22	56291 (23%)	96400 (39%)	19206 (8%)
Naproxen	-43.72	0.42	-0.35	1	0	-0.34	14738 (6%)	138633 (55%)	4298 (2%)
Piroxicam	-32.21	0.54	-0.19	3	1	-0.18	73815 (30%)	33269 (13%)	28731 (11%)
Resveratrol	-35.4	0.42	-0.17	0	1	-0.14	56672 (23%)	138543 (55%)	36765 (15%)
SC558	-38.26	0.75	-0.57	6	2	-0.51	40394 (16%)	6558 (3%)	215 (0%)

2. *Rscore* for DHFR

The *Rscore* analysis was repeated with *E. coli* dihydrofolate reductase (DHFR) [134]. In our study, we used 9 known inhibitors with nanomolar IC_{50} 's for DHFR. The receptor site was based on the crystal structure of 1RX3, complexed with methotrexate and NADP (the latter was included in the receptor definition used for docking). Only 7 of the 9 chosen inhibitors docked to the 1RX3 active site. These 7 inhibitors are shown in Figure 36.

The weights obtained for these 7 *DHFR* inhibitors using the linear programming formulation were $w_1 = 0.8$, $w_2 = -0.15$ and $w_3 = 0.35$. Table XIX shows the *DOCK* scores of these 7 inhibitors against the 9 decoy sites and the target receptor. All the 7 inhibitors bind with a positive score to the *ALR* receptor and inhibitor 446245 binds to the fewest number of decoy sites with a negative score (2/9). Inhibitor 22302034 binds to most of the receptor sites with larger than average *DOCK* scores, but the largest of these is with the *InhA* receptor (-205.69).

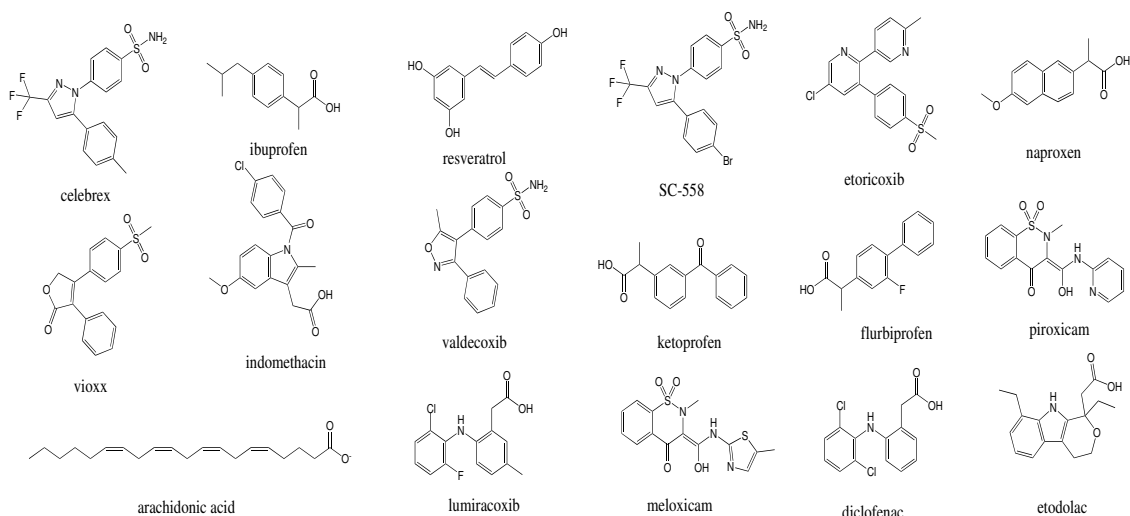


Fig. 34. Known COX-2 inhibitors used in this study

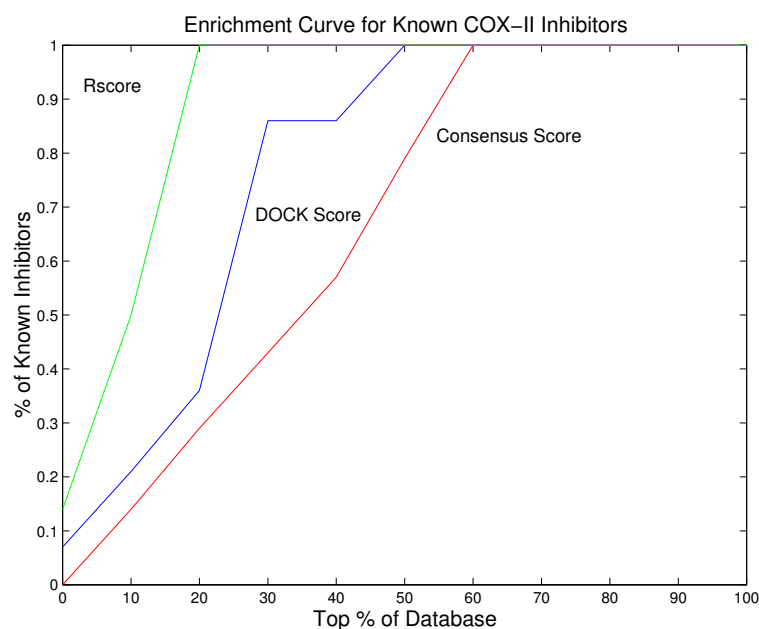


Fig. 35. The enrichment curves for COX-II based on the three different scores explored in this study. This graph shows that *Rscore* significantly increases the enrichment in comparison to both *DOCK* score as well as the consensus score from *Sybyl*

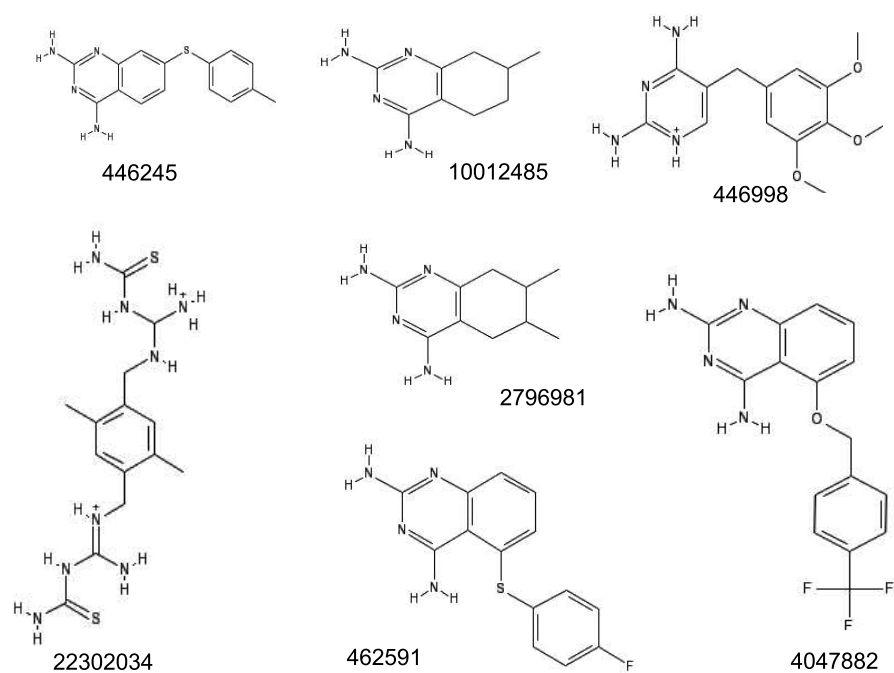


Fig. 36. Known DHFR inhibitors used in this study

Table XIX.: *DOCK* scores of known DHFR inhibitors across various receptors

Inhibitor	Active Site									
	ALR	AroD	BioA	DXR	InhA	MS	PanC	PfENR	PGDH	DHFR
10012485	13.93	-14.90	-29.30	-24.32	-27.02	-31.36	-27.30	-26.63	-11.58	-27.34
22302034	144.29	2.91	-70.30	-72.16	-205.69	-74.66		-52.04	-37.95	-62.19
2796981	463.88	-9.74	-28.89	-20.27	-27.41	-24.51	-28.92	-28.91	-9.78	-22.61
4047882	281.34	-28.43	-44.74	-41.92	-50.42	-46.34		-42.3	-20.50	-36.41
446245	37884	1048	897.4	-34.77	38.48	48.68		-32.03	109.6	-34.74
446998	51.55	18.71	-45.94	-49.67	-110.01	-55.86	-44.81	-43.1	-27.6	-42.98
462591	148347		-28.7	1137	-43.54	-16.06	140849	-30.54	135801	-36.33

Table XX shows the IC₅₀ values for the 7 inhibitors and the values of δ , π , ϕ . The rankings using the *DOCK* score, *Rscore* and consensus score are also tabulated in Table XX. While *DOCK* ranks only one of the known inhibitors near the top, and all the others around 100,000, *Rscore* ranks all the known inhibitors at approximately 10,000 or below (out of 250,000), and 3 within the top 100. The enrichment curve is shown in Figure 37.

Table XX.: Comparison of *Rscore* to *DOCK* score and consensus score for DHFR in virtual screen against ChemBridge library consisting of 250,000 compounds

Inhibitor (Pubchem CID)	IC ₅₀ (nM)	μ	δ	π	ϕ	<i>Rscore</i>	<i>DOCK</i> Rank	Consensus Score Rank	<i>Rscore</i> Rank
10012485	1.1×10^4	0.78	0.162	1	3	1.15	133319 (53%)	180669 (72%)	34464 (14%)
22302034	109	0.24	-0.243	2	4	1.11	5 (0%)	35994 (14%)	29012 (12%)
2796981	790	0.86	0.112	1	3	1.12	136967 (55%)	184852 (74%)	29473 (12%)
4047882	660	0.58	0.05	1	4	1.46	89770 (36%)	126946 (51%)	69511 (28%)
446245	310	0.93	-0.196	4	4	0.89	103648 (41%)	17843 (7%)	9607 (4%)
446998	18	0.41	-0.227	2	3	0.74	25583 (10%)	106949 (43%)	3381 (1%)
462591	400	0.98	-0.337	4	4	0.79	90493 (36%)	49973 (20%)	4734 (2%)

3. *Rscore* Results for Malate Synthase and Experimental Validation

Rscore ranking was applied to the results of virtual screen runs for malate synthase (MS) from *M. tuberculosis*. An unpublished crystal structure of MS complexed with a novel inhibitor (referred to as Compound **A** in this chapter) solved in our lab was used

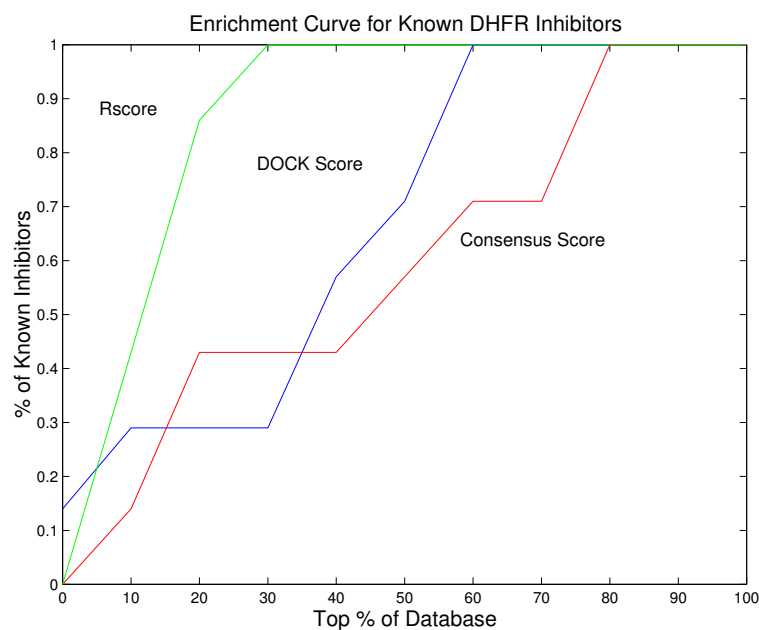


Fig. 37. The enrichment curves for DHFR based on the three different scores explored in this study. This graph shows that *Rscore* significantly increases the enrichment in comparison to both *DOCK* score as well as the consensus score from *Sybyl*

as the receptor for docking. This structure showed conformational changes compared to the known MS structure 2GQ3 complexed with CoA and malate[2].

The list of known inhibitors for MS included oxalate and parabanic acid (identified by Smith *et.al.* [117]) and three other inhibitors (henceforth referred to as Compounds **A**, **B** and **C**) that were identified by our collaborators. The weights obtained for these 5 MS inhibitors using the linear programming formulation were $w_1 = 0.978$, $w_2 = 0.02$ and $w_3 = 0.002$. Table XXI shows the dock scores of these known inhibitors against all the decoy sites and against the structure of MS complexed to Compound **A**. In this study COX-2 was used to replace MS as one of our decoy active sites. Oxalate binds to all the decoy sites with a negative score (small charged compound) whereas Compound **B** binds with a negative score to fewer decoy sites (3/9). Compound **B** is a larger molecule and this difference in size accounts for the fewer number of interactions with the decoy sites. Compound **A** which is close analog of Compound **B** also binds to fewer decoy sites.

Table XXI.: *DOCK* scores of known MS inhibitors across various receptors

Inhibitor	Active Site									
	ALR	AroD	BioA	COX-2	DXR	InhA	PanC	PfENR	PGDH	MS
oxalate	-36.39	-37.75	-26.61	-32.76	-31.13	9.38	-24.4	-27.79	-18.14	-50.01
parabanic acid	-22.93	-21.37	-23.14	-20.44	-21.7	-23.38	-22.84	-20.32	-20.11	-27.31
Compound A		-33.24	-50.35		-4.24	-74.53	-48.34			-45.17
Compound B		111			100	-34.05	-16.14	-46.47	186	-22.19
Compound C	-31.01	-40.67	-39.32	-42.45	-43.98	-16.42	-33.86	-38.03	-19.27	-64.67

Table XXII shows the ranking of these known inhibitors using the three different scoring schemes. Most of the inhibitors show an increase in ranking based on *Rscore* as compared to the ranking based on *DOCK* score by greater than 20%. The most dramatic increases are for parabanic acid and Compound **B** (37%).

Table XXII.: Comparison of *Rscore* to DOCK Score and consensus score for MS in virtual screen against ChemBridge library consisting of 250,000 compounds

Inhibitor	DOCK Score	μ	δ	π	ϕ	<i>Rscore</i>	DOCK Rank	Consensus Score Rank	<i>Rscore</i> Rank
oxalate	-50.01	0.53	-0.48	1	0	-0.45	37484 (15%)	135523 (54%)	32 (0%)
parabanic acid	-27.31	0.6	-0.19	0	0	-0.193	102811 (41%)	135519 (54%)	9651 (4%)
Compound A	-45.17	0.33	-0.22	0	4	-0.21	57547 (23%)	13045 (5%)	11114 (4%)
Compound B	-22.19	0.67	-0.23	3	3	-0.16	108181 (43%)	13025 (5%)	15002 (6%)
Compound C	-64.67	0.44	-0.44	0	0	-0.43	2460 (1%)	132720 (53%)	55 (0%)

The top 50 compounds as ranked by *Rscore* were manually analyzed and novel scaffolds were found. 16 compounds from the top 50 docked to the active site making interactions that have been observed crystallographically with known malate synthase ligands. At the same time, new interactions in the docked conformations of these top 50 compounds were suggested, providing novel ideas for inhibitor design. Using the assay described above, 4 out of these 16 compounds showed inhibition in the 100 μ M range and the inhibition values are listed in Table XXIII. These 4 compounds are shown in Figure 38. The *Tanimoto* similarity of these 4 compounds was computed against the five known inhibitors used in this study and these values were no greater than 0.3 between any of these compounds. This shows that in addition to finding new inhibitors, we have also found novel scaffolds of interaction (a primary motivation for virtual screening).

Table XXIII.: % Inhibition for novel inhibitors identified by *Rscore* ranking

Inhibitor	% Inhibition at 100 μ M
Compound 1	67%
Compound 2	41%
Compound 3	40%
Compound 4	24%

The hit rate of 4/16 is highly encouraging for a virtual screen run and efforts are

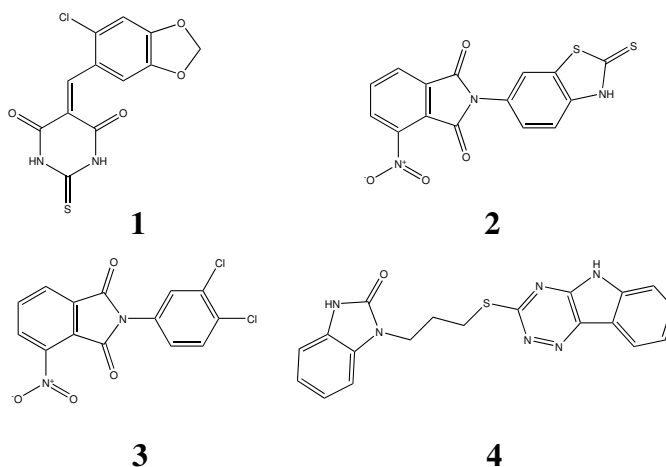


Fig. 38. The four inhibitors identified by *Rscore* for malate synthase. These four compounds have novel chemical scaffolds when compared to the previously known inhibitors

ongoing to purchase and test other compounds ranked highly by *Rscore*. The virtual screen results against the 2GQ3 (a relatively smaller and closed form) active site have also been reranked using *Rscore* and compounds from this set are also being tested in the laboratory for inhibition activity.

4. Promiscuous Virtual Screen Compounds from a Study of ChemBridge Library

Promiscuous small molecules in an HTS experiment are compounds that show high inhibition activity with a large number of diverse receptor sites [108] (these are distinct from aggregators [36]). A similar definition can be developed for promiscuous virtual screen compounds based on their docking scores against various receptors. If a compound is consistently ranked at the top of multiple virtual screen runs (large negative *DOCK* score against many receptors), then it can be assumed that it is making non-specific interactions with the receptors and the high scores reflect the bias of the scoring function. The *DOCK* score takes into account the electrostatic interactions

and the van der Waals interactions of the compound with the receptor. Therefore, the expectation is for large and charged compounds to be consistently ranked higher than other compounds in the library. In order to test this assumption, the docking scores of the compounds in the ChemBridge library across 10 diverse receptors were analyzed.

The compounds were sorted based on average rank, corrected for the number of decoy sites. Since some compounds only docked successfully to a few (less than 5) receptors, it is not as difficult to achieve a better (low) average relative rank by chance. In order to correct for this, we define a p-value that reflects the likelihood of getting a given average rank or better, compared to picking randomly from the same number of uniform distributions. The probability density that the average relative rank is r for k distributions is given by

$$P(x_1 + \dots + x_k = r) \propto \frac{1}{2(n-1)!} \sum_{i=0}^k (-1)^i \binom{k}{i} (r-i)^{k-1} \text{sgn}(r-i) \quad (6.12)$$

The cumulative probability $P(x_1 + \dots + x_k > r)$ of a given average rank of r or better was estimated by Monte Carlo sampling. The compounds were sorted by p-value and 100 compounds that docked to a maximum number of decoy sites were chosen for further analysis. Most ($> 90\%$) of the compounds docked to 6 or more sites with an average rank from 0 to 10%.

Figure 39 shows the superposition of the distributions of molecular weight for the top 100 virtually promiscuous compounds identified in our study with p-value $< 5 \times 10^{-5}$ and those of all the compounds in the ChemBridge library. This figure shows that approximately 85% of these compounds have a molecular weight between 245 to 350. 25% of the ChemBridge database has compounds with greater molecular weights than the selected promiscuous compounds. Therefore, we analyzed the number of

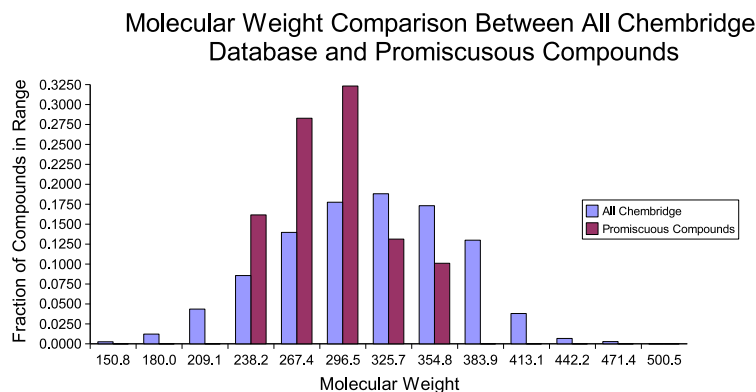


Fig. 39. The majority of the top 100 virtually promiscuous compounds have a molecular weight greater than 275

rotatable bonds in these top 100 molecules and the ChemBridge database and these are plotted in Figure 40. This figure shows that 90% of the top 100 molecules had 6 or greater rotatable bonds while approximately 50% of the ChemBridge database has fewer than 6 rotatable bonds. The virtually promiscuous compounds seem to show a tendency towards larger rotatable bonds. This high flexibility in these compounds make it more likely that some conformation of the molecules would be able to make non-specific interactions with multiple active sites.

Figure 41 shows the superimposition of the distributions of net charge across top 100 virtually promiscuous compounds and those of all the compounds in the ChemBridge library. This figure shows that 70% of the promiscuous molecules were neutrally charged contrary to expectations. A deeper study of these compounds shows that while the overall charge for these compounds is zero, they have positively charged as well as negatively charged components (zwitterionic compounds). This observation led us to analyze the polar desolvation energy (obtained from the ZINC database) of these top 100 compounds and this is plotted in Figure 42. This figure shows that the polar desolvation penalty for 40% of the promiscuous compounds

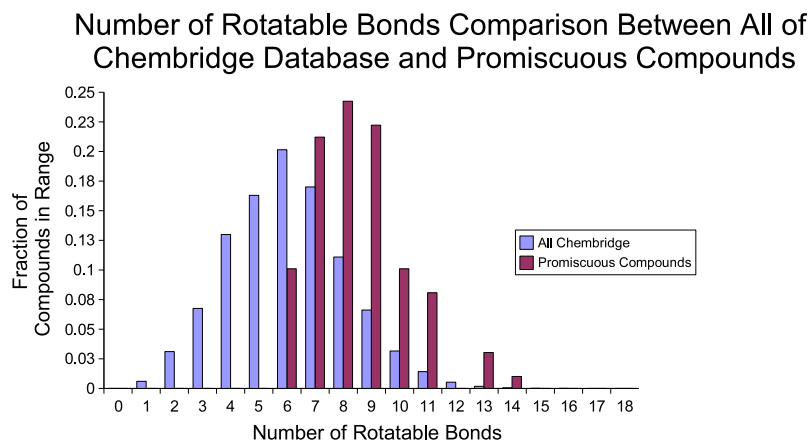


Fig. 40. The majority of the top 100 virtually promiscuous compounds have greater than 7 rotatable bonds

is lower than -30 and in this range there are fewer compounds from the rest of the ChemBridge library. The compounds with higher desolvation penalties (lower than -30) are those that have greater localized charge since localized charges lead to less favorable desolvation energies as compared to ligands with delocalized charges. [18]. Figure 43 shows a subset of these top 100 compounds.

This analysis of virtually promiscuous compounds refines our earlier assumptions of the bias of *DOCK* scores to large and charged compounds. Our results show that the molecular weight of the compound is not as relevant as the flexibility of the compound (as dictated by the number of non-rotatable bonds). Also, looking at the overall charge of a molecule might be deceptive and some of the neutral molecules in the library do in fact have charge cancellation. The presence of localized charges might be more accurately captured by the polar desolvation energy term.

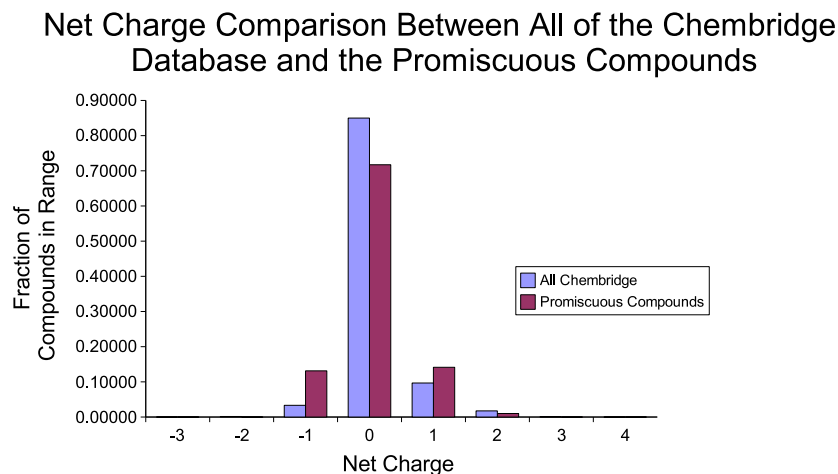


Fig. 41. 70% of the top 100 virtually promiscuous compounds are either positively or negatively charged. The overall charge for the remaining compounds is zero but they contain both positively and negatively charged components.

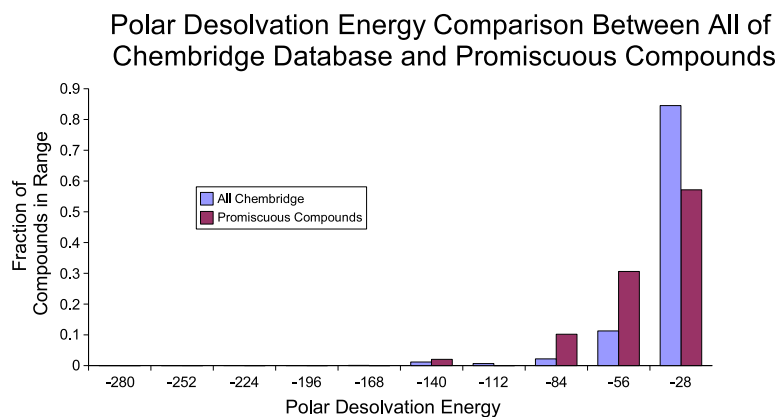


Fig. 42. The distribution of polar desolvation energy for the ChemBridge database and the top 100 virtually promiscuous compounds

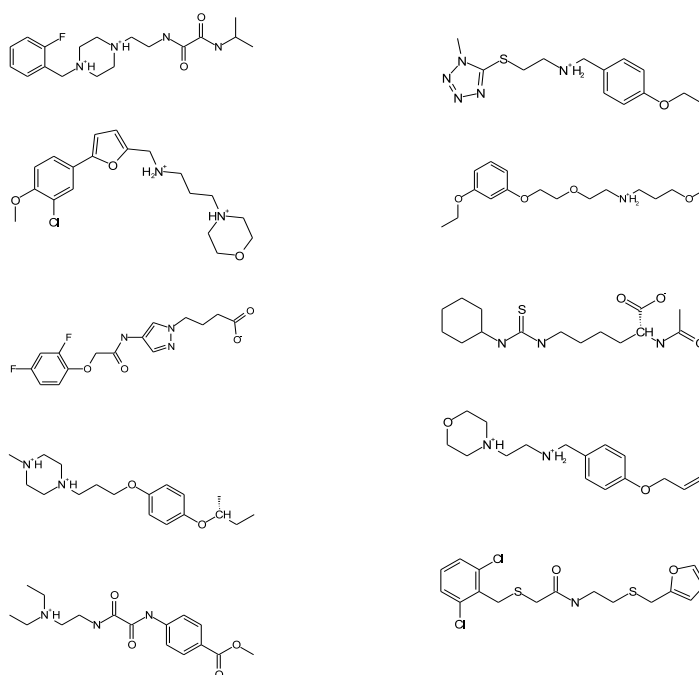


Fig. 43. This figure shows some of the promiscuous virtual screen compounds identified by our analysis of the ChemBridge library across 10 diverse receptor sites

D. Discussion

Rscore helps remove biases in the scoring function (*e.g.* preference toward large and charged compounds) and thereby promotes diversity within the top ranked compounds. Additionally, despite its use of known inhibitors in its analysis, it does not necessarily bias the results towards the scaffold of known inhibitors. It only seeks to mimic the interaction profile of the known inhibitors across the decoy sites (*i.e.* those that interact favorably to the target receptor and unfavorably to the decoy sites). Therefore it retains the diversity of selections from the database.

In this study, we have assumed all of the compounds in the library are non-inhibitors. Examining the chemical similarity between the known inhibitors and the compounds in the small-molecule library could be used to identify compounds with similar chemical profiles and these compounds can then be additionally considered as positive examples in the linear programming formulation. Since the formulation of *Rscore* depends on known inhibitors, any increase in the number of known inhibitors used in training will improve the reliability of the weights obtained thereby increasing the reliability of *Rscore*.

Essential to the definition of *Rscore* is the docking of small-molecules to the decoy active sites. While the process of docking 250,000 compounds to decoy active sites is time-consuming, these jobs have to be run only once and the results can be used for normalizing subsequent virtual screen runs. The number of decoy sites is variable and a larger number of sites can only increase the accuracy of the approach. The computation of weights using linear programming is very simple and fast.

E. Conclusions

In this chapter, we have experimentally validated a novel quantitative approach to increase the recall and enrichment in a virtual screen. This methodology increases the rank of some of the known inhibitors by almost 20%. This improvement in ranking additionally translates to a high hit ratio (ability to identify inhibitors) for virtual screens, making it a simple, yet powerful tool to re-rank the results of a virtual screen without having to modify the scoring function.

The definition of virtually promiscuous small molecules provides a detailed analysis of the bias of *DOCK* (or other scoring functions using a similar analysis). It also offers a way to reduce the computational burden on large virtual screens by pre-identifying these compounds in databases and removing them.

CHAPTER VII

CONCLUSIONS AND FUTURE WORK

In this dissertation, a novel approach to active site analysis was developed. To our knowledge, this study is the most comprehensive study of protein-ligand interactions to date. No previous methodology has been tested on as many diverse proteins bound to diverse ligands. The approach in this dissertation uses a fragment-based analysis to account for the flexibility observed in protein-ligand interactions and a granular feature-based representation of the active site. This two-stage process addresses two significant issues with traditional methods of functional annotation.

The feature-based analysis presented in this study stepped away from the traditional sequence homology, fold family analyses and global features for active site descriptions. It utilized more granular stereochemical features to capture the interaction patterns within the active site in greater detail and with greater fidelity, ultimately yielding greater classification accuracy. The position-dependent features performed much better than the localized features by correctly categorizing more fragment classes and an accuracy of 84%. Additionally, the feature-based analysis, most often, successfully identified the individual fragments within each ligand thereby increasing the probability of identifying the entire ligand. The *mrf* analysis provided a strong mathematical foundation for the ligand fragment combination by considering both the fragment classifications and the statistical model describing the distance relationships between ligand fragments. This procedure was able to successfully identify the true ligand within the top 10 results in almost all test cases. The overall procedure was tested on 3 proteins with very low sequence and structural similarity to other proteins in the *PDB* (a challenge for traditional methods) and in each of these cases, the approach presented in this dissertation, successfully identified the cognate

ligand.

1. Future Work

The high accuracy of this fragment-based analysis is encouraging and several avenues exist to increase the relevance and efficiency of this approach. The database used in this study was based on the current information available regarding protein-ligand complexes. The sparsity of this information restricted the size of our database to fewer variety of ligands and also fewer number of examples for some ligands. Efforts to automate literature surveys in order to identify other protein-ligand interactions within the active site and making this data publicly available will greatly benefit future active site analysis efforts.

The current geometric and electrostatic sector features have proven to be powerful in distinguishing between 441 different fragment classes, but as the size of the database increases, these features might need to be refined. In the present electrostatic analysis, the effects of metal ions to active site chemistry is taken into account. However, the effect of other cofactors within the active site is not considered. Very often, the binding of ligands in the active site is coordinated and very dependent on the other cofactors and including a chemical analysis of these cofactors will increase the accuracy of the electrostatic analysis. Similarly, information regarding the protonation states of various protein residues could also be incorporated into the electrostatic analysis. This information is often speculative/unavailable. The methodology developed in this dissertation makes it very easy to incorporate these new features. The most computationally intensive task is to define the active site surfaces and additional feature calculations are not as intensive.

Another highly debated issue in active site analysis is the consideration of water atoms in/near the active site. Very often, ligands co-ordinate water molecules in order

to make highly specific electrostatic interactions. In this study, these molecules were excluded since the identification of essential waters is often difficult and once again this information is not easily available.

Each of these avenues holds the potential to make this algorithm more powerful and relevant for the analysis of diverse active sites and are currently being pursued.

REFERENCES

- [1] R. Abagyan, M. Totrov and D Kuznetsov. “ICM - A new method for protein modelling and design: Applications to docking and structure prediction from the distorted native conformation,” *J. Comput. Chem.*, vol. 15, pp. 488–506, 1994.
- [2] D.M. Anstrom and S.J. Remington. “The product complex of *M. tuberculosis* malate synthase revisited,” *Protein Sci.*, vol. 15, pp. 2002–2007, 2006.
- [3] R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M.D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N.J. Mulder, T.M. Oinn, M. Pagni, F. Servant, C.J. Sigrist and E.M. Zdobnov. “The InterPro database, an integrated documentation resource for protein families, domains and functional sites,” *Nucleic Acids Res.*, vol. 29, no. 1, pp. 37–40, 2001.
- [4] P. Artymiuk, A. Poirrette, H. Grindley, D. Rice and P. Willett. “A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures,” *J. Mol. Biol.*, vol. 243, pp. 327–344, 1994.
- [5] T.K. Attwood, M.E. Beck, A.J. Bleasby and D.J. Parry-Smith. “PRINTS - A database of protein motif fingerprints”, *Nucleic Acids Res.*, vol. 22, no. 17, pp. 3590–3596, 1994.
- [6] S.C. Bagley, L. Wei, C. Cheng and R.B. Altman. “Characterizing oriented protein structural sites using biochemical properties,” in *Proc. 3rd Intl. Conf. on Intelligent Systems for Mol. Biol.*, vol. pp. 12–20, 1995.

- [7] S.C. Bagley and R.B. Altman. “Characterizing the microenvironment surrounding protein sites,” *Protein Science*, vol. 4, no. 4, pp. 622–635, 1995.
- [8] J.A. Barker and J.M. Thornton. “An algorithm for constraint-based structural template matching: Application to 3D templates with statistical analysis,” *Bioinformatics*, vol. 19, pp. 1644–1649, 2003.
- [9] G.J. Bartlett, C.J. Porter, N. Borkakoti and J.M. Thornton. “Analysis of catalytic residues in enzyme active sites,” *J. Mol. Biol.*, vol. 324, pp. 105–121, 2002.
- [10] J. Benz, H. Trachsel and U. Baumann. “Crystal structure of the ATPase domain of translation initiation factor 4A from *Saccharomyces cerevisiae*- The prototype of the DEAD box protein family,” *Structure*, vol. 7, pp. 671–679, 1999.
- [11] C. Berezin, F. Glaser, J. Rosenberg, I. Paz, T. Pupko, P. Fariselli, R. Casadio, and N. Ben-Tal. “ConSeq: The identification of functionally and structurally important residues in protein sequences,” *Bioinformatics*, vol. 20, no. 8, pp. 1322–1324, 2004.
- [12] J. Besag. “Spatial interaction and the statistical analysis of lattice systems,” *Journal of the Royal Statistical Society, Series B*, vol. 36, pp. 192–236, 1974.
- [13] J.R. Beveridge, K. She, B. Draper, and G.H. Givens. “A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition”, *IEEE Conference on Pattern Recognition and Machine Intelligence*, pp. 535–542, 2001.
- [14] R.S. Bohacek, and C. McMartin. “Multiple highly diverse structures complementary to enzyme binding sites: Results of extensive application of a *de nova* design method incorporating combinatorial growth,” *J. Am. Chem. Soc.*, vol. 116, pp. 5560–5571, 1994.

- [15] H.J. Bohm. “The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure,” *J. Comput. Aided. Mol. Design.*, vol. 8, no. 3, pp. 243–256, 1994.
- [16] P. Bork, C. Sander and A. Valencia. “An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins”, *PNAS*, vol. 89, no. 16, pp. 7290–7294, 1992.
- [17] J. Bostrom, J.R. Greenwood, J. Gottfires. “Assessing the performance of OMEGA with respect to retrieving bioactive conformations,” *J. Mol. Graphics and Mod.*, vol. 21, pp. 449–462, 2003.
- [18] R. Brenka, S.W. Vetterb, S.E. Boycea, D.B. Goodinb, and B.K. Shoichet. “Probing molecular docking in a charged model binding site,” *J. Mol. Biol.*, vol. 357, no. 5, pp. 1449–1470, 2006.
- [19] P. Burkhard, U. Hommel, M. Sanner and M.D. Walkinshaw. “The discovery of steroids and other novel FKBP inhibitors using a molecular docking program,” *J Mol Biol.*, vol. 287, no. 5, pp. 853–858, 1999.
- [20] E.P. Carpenter, A.R. Hawkins, J.W. Frost and K.A. Brown. “Structure of dehydroquinase synthase reveals an active site capable of multistep catalysis,” *Nature*, vol. 394, pp. 299–302, 1998.
- [21] V. Cerny. “A thermodynamical approach to the travelling salesman problem: An efficient simulation algorithm,” *Journal of Optimization Theory and Applications*, vol. 45, pp. 41–51, 1985.
- [22] P.S. Charifson, J.J. Corkery, M.A. Murcko and W.P. Walters. “Consensus scoring: A method for obtaining improved hit rates from docking databases of three-

- dimensional structures into proteins,” *J Med Chem.*, vol. 42, no. 25, pp. 5100–5109, 1999.
- [23] P. Chavatte, S. Yous, C. Marot, N. Baurin and D. Lesieur. “Three-dimensional quantitative structure-activity relationships of cyclo-oxygenase-2 no. COX-2) inhibitors: A comparative molecular field analysis,” *J Med Chem.*, vol. 44, no. 20, pp. 3223–3230, 2001.
- [24] M. Connolly. “Analytical molecular surface calculation,” *J. Appl. Cryst.*, vol. 16, pp. 548–558, 1983.
- [25] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell and P.A. Kollman. “A second generation force field for the simulation of proteins, nucleic acids, and organic molecules,” *J. Am. Chem. Soc.*, vol. 117, pp. 5179–5197, 1985.
- [26] M.O. Dayhoff, R.M. Schwartz and B.C. Orcutt. “A model of evolutionary change in proteins,” in Dayhoff, M. O., Ed., *Atlas of Protein Sequence and Structure*, vol. 5, National Biomedical Research Foundation, Washington, DC, pp. 345–352, 1978.
- [27] K.A. Denessiouk and M.S. Johnson. “When fold is not important: A common structural framework for adenine and AMP binding in 12 unrelated protein families,” *Proteins:Structure, Function and Genetics*, vol. 38, pp. 310–326, 2000.
- [28] K.A. Denessiouk, V. Rantanen and M.S. Johnson. “Adenine recognition: A motif present in ATP, CoA, NAD, NADP and FAD dependent proteins,” *Proteins:Structure, Function and Genetics*, vol. 44, pp. 282–291, 2001.
- [29] K.A. Denessiouk and M.S. Johnson. “Acceptor-donor-acceptor motifs recognize the Watson-Crick, Hoogsteen and sugar-donor-acceptor-donor edges of adenine

- and adenosine-containing ligands,” *J. Mol. Biol.*, vol. 333, no. 5, pp. 1025–1043, 2003.
- [30] S. Dey and J.C. Sacchettini. “Crystal structure of PLP bound 7,8-diaminopelargonic acid synthase in *Mycobacterium tuberculosis*,” *To be Published*.
- [31] S. Dey, G.A. Grant and J.C. Sacchettini. “Crystal structure of *Mycobacterium tuberculosis* D-3-phosphoglycerate dehydrogenase: extreme asymmetry in a tetramer of identical subunits,” *J.Biol.Chem.*, vol. 280, pp. 14892–14899, 2005.
- [32] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*, New York, 1973.
- [33] D.A. Erlanson, J.A. Wells and A.C. Braisted. “TETHERING: Fragment-based drug discovery,” *Annu. Rev. Biophys. Biomol. Struct.*, vol. 33, pp. 199–223, 2004.
- [34] G. Ermondia, G. Carona, R. Lawrence and D. Longo. “Docking studies on NSAID/COX-2 isozyme complexes using Contact Statistics analysis,” *Jol. of Comp. Aided Mol. Des.*, vol. 18, pp. 683–696, 2004.
- [35] J.L. Fauchere and V. Pliska. “Hydrophobic parameters- π of amino-acid side-chains from the partitioning of N-acetyl-aminoacid amides,” *Eur. J. Med. Chem.*, vol. 18, pp. 369–375, 1983.
- [36] B.Y. Feng, A. Shelat, T.N. Doman, R.K. Guy and B.K. Shoichet. “High-throughput assays for promiscuous inhibitors,” *Nature Chemical Biology*, vol. 1, pp. 146–148, 2005.
- [37] J.S. Fetrow and J. Skolnick. “Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application

- to glutaredoxins/thioredoxins and T_1 ribonucleases,” *J. Mol. Biol.*, vol. 281, pp. 949–968, 1998.
- [38] A.V. Filikov, V. Mohan, T.A. Vickers, R.H. Griffey, P.D. Cook, R.A. Abagyan and T.L. James. “Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR,” *J Comput Aided Mol Des.*, vol. 14, no. 6, pp. 593–610, 2000.
- [39] K.M. Flaherty, D.B. McKay, W. Kabsch and K.C. Holmes. “Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70 kDa heat shock cognate protein,” *Proc. Nat. Acad. Sci.*, vol. 88, pp. 5041–5045, 1991.
- [40] I. Friedberg and H. Margalit. “Persistently conserved positions in structurally similar, sequence dissimilar proteins: Roles in preserving protein fold and function,” *Protein Sci.*, vol. 11, no. 2, pp. 350–360, 2002.
- [41] F. Glaser, R.J. Morris, R.J. Najmanovich, R.A. Laskowski. and J.M. Thornton. “A method for localizing ligand binding pockets in protein structures,” *Proteins:Structure, Function and Bioinformatics*, vol. 62, pp. 479–488, 2006.
- [42] R. Ginton, J. Giampapa and K. Sycara. “A Markov random field model of context for high-level information fusion,” in *Proc. Ninth International Conference on Information Fusion*, pp. 1–8, 2006.
- [43] J.W. Godden, F. Stahura and J. Bajorath. “Evaluation of docking strategies for virtual screening of compound databases: cAMP-dependent serine/threonine kinase as an example,” *J Mol Graph Model.*, vol. 16, no. 3, pp. 139–143, 1999.
- [44] J.C. Gordon, J.B. Myers, T. Folta, V. Shoja, L.S. Heath, A. Onufriev. “H++: A server for estimating pKas and adding missing hydrogens to macromolecules,” *Nucleic Acids Res.*, vol. 33, pp. W368–371, 2005.

- [45] D.G. Gourley, A.K. Shrive, I. Polikarpov, T. Krell, J.R. Coggins, A.R. Hawkins, N.W. Isaacs and L. Sawyer. “The two types of 3-dehydroquinase have distinct structures but catalyze the same overall reaction,” *Nature Structural Biology*, vol. 6, pp. 521–525, 1999.
- [46] A. Gutteridge, G.J. Bartlett and J.M. Thornton. “Using a neural network and spatial clustering to predict the location of active sites in enzymes,” *J. Mol. Biol.*, vol. 330, pp. 719–734, 2003.
- [47] J.Y. Ha, J.Y. Min, S.K. Lee, H.S. Kim, D.J. Kim, K.H. Kim, H.H. Lee, H.K. Kim, H.J. Yoon and S.W. Suh. “Crystal structure of 2-nitropropane dioxygenase complexed with FMN and substrate,” *Journal of Biological Chemistry*, vol. 281, no. 27, pp. 18660–18667, 2006.
- [48] S. Henikoff and J.G. Henikoff. “Amino acid substitution matrices from protein blocks,” *Proc. Natl Acad. Sci.*, vol. 89, no. 22, pp. 10915–10919, 1992.
- [49] L.M. Henriksson, T. Unge, J. Carlsson, J. Aqvist, S.L. Mowbray and T.A. Jones. “Structures of *Mycobacterium tuberculosis* 1-Deoxy-D-Xylulose-5-Phosphate reductoisomerase provide new insights into catalysis,” *J.Biol.Chem.*, vol. 282, pp. 19905, 2007.
- [50] J.C. Hermann, R. Marti-Arbona, A.A. Fedorov, E. Fedorov, S.C. Almo, B.K. Shoichet and F.M. Raushel. “Structure-based activity prediction for an enzyme of unknown function,” *Nature*, vol. 448, pp. 775–779, 2007.
- [51] S.A. Hindle, M. Rarey, C. Buning and T. Lengau. “Flexible docking under pharmacophore type constraints,” *J. Computer Aided Mol. Des.*, vol. 16, no. 2, pp. 129–149, 2002.

- [52] L. Holm and C. Sander. “Enlarged representative set of protein structures,” *Protein Science*, vol. 3, pp. 522–524, 1994.
- [53] L. Holm and C. Sander. “Mapping the protein universe,” *Science*, vol. 273, pp. 595–602, 1996.
- [54] L. Holm and J. Park. “DaliLite workbench for protein structure comparison,” *Bioinformatics*, vol. 16, pp. 566–567, 1999.
- [55] B. Honig, and A. Nicholls. “Classical electrostatics in biology and chemistry,” *Science*, vol. 268, no. 5214, pp. 1144–1149, 1995.
- [56] S.J. Hubbard and J.M. Thornton. “NACCESS,” Department of Biochemistry and Molecular Biology, University College London., 1993.
- [57] 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.*, vol. 267, pp. 727–748, 1997.
- [58] S. Jones and J.M. Thornton. “Analysis of protein-protein interaction sites using patch analysis,” *J. Mol. Biol.*, vol. 272, pp. 121–132, 1997.
- [59] W. Kabsch and C. Sander. “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features,” *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.
- [60] A. Kahraman, R. J. Morris, R.A. Laskowski and J.M. Thornton. “Shape variation in protein binding pockets and their ligands,” *J Mol Biol.*, vol. 368, no. 1, pp. 283–301, 2007.
- [61] A. Kasuya and J.M. Thornton. “Three-dimensional structural analysis of PROSITE patterns,” *J. Mol. Biol.*, vol. 286, pp. 1673–1691, 1999.

- [62] G.W. Kauffman and P.C. Jurs. “QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors,” *J Chem Inf Comput Sci.*, vol. 41, no. 6, pp. 1553–1560, 2001.
- [63] M.S. Kimber and E.F. Pai. “The active site architecture of *Pisum sativum* β -carbonic anhydrase is a mirror image of that of β -carbonic anhydrases,” *EMBO Journal*, vol. 19, pp. 1407–1418, 2000.
- [64] S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi. “Optimization by simulated annealing,” *Science*, vol. 220 no. 4598, pp. 671–680, 1983.
- [65] T. Kobori, H. Sasaki, W.C. Lee, S. Zenno, K. Saigo, M.E. Murphy and M. Tanokura. “Structure and site-directed mutagenesis of a flavoprotein from *Escherichia coli* that reduces nitrocompounds: Alteration of pyridine nucleotide binding by a single amino acid substitution,” *J. Mol. Biol.*, vol. 276, no. 4, pp. 2816–2823, 2001.
- [66] H. Kubinyi. “Drug research: Myths, hype, and reality,” *Nature Reviews Drug Discovery*, vol. 2, 665–668, 2003.
- [67] D. Kuhn, N. Weskamp, E. Hullermeier and G. Klebe. “Functional classification of protein kinase binding sites using Cavbase,” *ChemMedChem*, vol. 2, no. 10, pp. 1432–1447, 2007.
- [68] R.G. Kurumbail, A.M. Stevens, J.K. Gierse, J.J. McDonald, R.A. Stegeman, J.Y. Pak, D. Gildehaus, J.M. Miyashiro, T.D. Penning, K. Seibert, P.C. Isakson and W.C. Stallings. “Structural basis for selective inhibition of cyclooxygenase-2 by anti-inflammatory agents,” *Nature*, vol. 384, pp. 644–648, 1996.

- [69] Y. Kuttner, V. Soboloev, A. Raskind and M. Edelman. “A consensus-binding structure for adenine at the atomic level permits searching for the ligand site in a wide spectrum of adenine-containing complexes,” *Proteins Struct. Func. Genet.*, vol. 52, pp. 400–411, 2003.
- [70] R.A. Laskowski, J.D. Watson and J.M. Thornton. “Protein function prediction using local 3D templates,” *J.Mol.Biol.*, vol. 351, pp. 614–626, 2005.
- [71] J.D. Lawson, E. Pate, I. Rayment and R.G. Yount. “Molecular dynamics analysis of structural factors influencing backdoor P_i release in myosin,” *Biophysics Journal*, vol. 86, no. 6, pp. 3794–3803, 2004.
- [72] P. LeMagueres, H. Im, J. Ebalunode, U. Strych, M.J. Benedik, J.M. Briggs, H. Kohn and K.L. Krause. “The 1.9Å crystal structure of alanine racemase from *Mycobacterium tuberculosis* contains a conserved entryway into the active site,” *Biochemistry*, vol. 44, pp. 1471–1481, 2005.
- [73] A.J. Li and R. Nussinov. “A set of van der Waals and colombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation and docking,” *Proteins: Structure, Function and Genetics*, vol. 32, pp. 11–127, 1998.
- [74] S.Z. Li. Markov Random Field modeling in image analysis. Kunii TL. Computer Science Workbench, New York Springer Verlag 2001.
- [75] J. Liang, H. Edelsbrunner and C. Woodward. “Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design,” *Protein Science*, vol. 7, pp. 1884–1897, 1998.
- [76] O. Lichtarge, H.R. Bourne and F.E. Cohen. “The evolutionary trace method

- defines the binding surfaces common to a protein family,” *J. Mol. Biol.* vol. 257, pp. 342–358, 1996.
- [77] P. Linder. “Dead-box proteins: A family affair active and passive players in RNP-remodeling,” *Nucleic Acids Research*, vol. 34, no. 15, pp. 4168–4180, 2006.
- [78] A.D. Mackarell. “Empirical force fields for biological macromolecules: Overview and issues,” *J. Comp. Chem*, vol. 25, pp. 1584–1604, 2004.
- [79] A.M. Martinez and A.C. Kak. “PCA versus LDA,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.
- [80] M.St. Maurice, P.E. Mera, M.P. Taranto, F. Sesma, J.C. Escalante-Semerena, I. Rayment. “Structural characterization of the active site of the PduO-type ATP:Co no. I)rrinoid adenosyltransferase from *Lactobacillus reuteri*,” *J. Biol. Chem*, vol. 282, no. 4, pp. 2596–2605, 2007.
- [81] E.C. Meng, B.K. Shoichet and I.D. Kuntz. “Automated docking with grid-based energy evaluation,” *J.Comp.Chem.*, vol. 13, pp. 505–524, 1992.
- [82] L.A. Mirny and E.I. Shakhnovich. “Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics and function,” *J. Mol. Biol.*, vol. 291, pp. 177–196, 1999.
- [83] S.L. Moodie, J.B. Mitchell and J.M. Thornton. “Protein recognition of adenylate: An example of a fuzzy recognition template,” *J. Mol. Biol.*, vol. 263, no. 3, pp. 486–500, 1996.
- [84] N. Moriarty. *Python-based Hierarchical ENvironment for Integrated Xtallography*, <http://www.phenix-online.org/documentation/elbow.htm>

- [85] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew and A.J. Olson. “Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function,” *J. Computational Chemistry*, vol. 19, pp. 1639–1662, 1998.
- [86] D.T. Moustakas, P.T. Lang, S. Pegg, E. Pettersen, I.D. Kuntz, N. Brooijmans and R.C. Rizzo. “Development and validation of a modular, extensible docking program: DOCK 5,” *J Comput Aided Mol Des.*, vol. 20 no. 10-11, pp. 601–619, 2006.
- [87] J.C. Mozziconacci, N. Baurin, L. Morin-Allory and C. Marot. “Automated docking of cox-2 selective inhibitors for virtual screening and combination with 2D-QSAR approach,” in *Proc. Eighth Electronic Computational Chemistry Conference*, 2002.
- [88] I. Muegge and Y.C. Martin. “A general and fast scoring function for protein-ligand interactions: A simplified potential approach,” *J. Med. Chem.*, vol. 42, pp. 791–804, 1999.
- [89] A.G. Murzin, S.E. Brenner, T. Hubbard and C. Chothia. “SCOP: A structural classification of proteins database for the investigation of sequences and structures,” *J. Mol. Biol.*, vol. 247, pp. 536–540, 1995.
- [90] N. Nagano, C.A. Orengo and J.M. Thornton. “One fold with many functions: The evolutionary relationships between TIM Barrel families based on their sequences, structures and functions,” *J Mol Biol.*, vol. 321, no. 5, pp. 741–765, 2002.
- [91] N. Nagano. “EzCatDB: The enzyme catalytic-mechanism database,” *Nucleic Acids Res.*, vol. 33, pp. 407–412, 2005.

- [92] A. Oda, K. Tsuchida, T. Takakura, N. Yamaotsu and S. Hirono. "Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes," *J. Chem. Inf. Model.*, vol. 46, no. 1, pp. 380–391, 2006.
- [93] T.J. O'Donnell. *gNOVA Scientific Software*, <http://www.gnova.com/docs/makefp.html>
- [94] OEChem, version 1.3.4, *OpenEye Scientific Software, Inc.*, Santa Fe, NM, USA, www.eyesopen.com, 2005.
- [95] S. Olgen, E. Akaho and D. Nebioglu. "Synthesis and receptor docking studies of N-substituted indole-2-carboxylic acid esters as a search for COX-2 selective enzyme inhibitors," *Eur J Med Chem.*, vol. 36, no. 9, pp. 747–770, 2001.
- [96] R. Pai, J.C. Sacchettini and T.R. Ioerger. "Analysis of protein-ligand interactions using localized stereochemical features," in *Proc. International Conference on Bioinformatics and Computational Biology no. BIOCAMP'08*, pp. 666–673, 2008.
- [97] R. Pai, J.C. Sacchettini and T.R. Ioerger. "Specificity normalization for identifying selective inhibitors in virtual screening," in *Proc. International Conference on Bioinformatics and Computational Biology no. BIOCAMP'08*, pp. 787–793, 2008.
- [98] A. Palomer, J. Pascual, M. Cabre, L. Borrás, G. Gonzalez, M. Aparici, A. Carabaza, F. Cabre, M.L. Garcia and D. Mauleon. "Structure-based design of cyclooxygenase-2 selectivity into Ketoprofen," *Bioorg Med Chem Lett.*, vol. 12, no. 4, pp. 533–537, 2002.
- [99] F. Pazos and M.J.E. Sternberg. "Automated prediction of protein function and

- detection of functional sites from structure,” *PNAS*, vol. 101, no. 41, pp. 14754–14759, 2004.
- [100] R. Perozzo, M. Kuo, A.S. Sidhu, J.T. Valiyaveetil, R. Bittman, W.R. Jacobs, D.A. Fidock and J.C. Sacchettini. “Structural elucidation of the specificity of the antibacterial agent Triclosan for malarial enoyl acyl carrier protein reductase,” *J.Biol.Chem.*, vol. 277, pp. 13106–13114, 2002.
- [101] C.T. Porter, G.J. Bartlett and J.M. Thornton. “The Catalytic Site Atlas: A resource of catalytic sites and residues identified in enzymes using structural data,” *Nucl. Acids. Res.*, vol. 32, pp. D129–D133, 2004.
- [102] R. Powers, J.C. Copeland, K. Germer, K.A. Mercier, V. Ramanathan and P. Revesz. “Comparison of protein active site structures for functional annotation of proteins and drug design,” *Proteins:Structure, Function and Bioinformatics*, 65, pp. 124–135, 2006.
- [103] M. Rarey, B. Kramer, T. Lengauer and G. Klebe. “A fast flexible docking method using an incremental construction algorithm,” *J Mol Biol.*, vol. 261, no. 3, pp. 470–489, 1999.
- [104] D.C. Rees, M. Congreve, C.W. Murray and R. Carr. “Fragment-based lead discovery,” *Nature Reviews Drug Discovery*, vol. 3, no. 8, pp. 660–672, 2004.
- [105] F.M. Richards. “Areas, volumes, packing and protein structure,” *Annual Review of Biophysics and Bioengineering*, vol. 6, pp. 151–176, 1977.
- [106] D. Riendeau, M.D. Percival, S. Boyce, C. Brideau, S. Charleson et al. “Biochemical and pharmacological profile of a tetrasubstituted furanone as a highly selective COX-2 inhibitor,” *Br J Pharmacol.*, vol. 121, no. 1, pp. 105–117, 1997.

- [107] D. Riendeau, M.D. Percival, C. Brideau, S. Charleson et al. "Etoricoxib no. MK-0663): Preclinical profile and comparison with other agents that selectively inhibit cyclooxygenase-2," *J Pharmacol Exp Ther.*, vol. 296, no. 2, pp. 558–566, 2001.
- [108] G.M. Rishton. "Reactive compounds and in vitro false positives in HTS," *Drug Discov. Today*, vol. 2, pp. 382–384, 1997.
- [109] D.A. Robinson, A.W. Roszak, M. Frederickson, C. Abell, J.R. Coggins and A.J. Laphorn. "Structural basis for selectivity of oxime based inhibitors towards type II dehydroquinase from *Mycobacterium tuberculosis*," *To be Published*.
- [110] D.A. Rozwarski, G.A. Grant, D.H. Barton, W.R. Jacobs, J.C. Sacchettini. "Modification of the NADH of the isoniazid target no. InhA) from *Mycobacterium tuberculosis*," *Science*, vol. 279, pp. 98–102, 1998.
- [111] L. Rychlewski, B. Zhang and A. Godzik. "Fold and function predictions for *Mycoplasma genitalium* proteins," *Fold Des.*, vol. 3, no. 4, pp. 229–238, 1998.
- [112] F.A. Sadjadi and E.L. Hall. "Three-dimensional moment invariants," *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 2, no. 2, pp. 127–136, 1980.
- [113] H.P. Shanahan and J.M. Thornton. "An examination of the conservation of surface patch polarity for proteins," *Bioinformatics*, vol. 20 no. 14, pp. 2197–2204, 2004.
- [114] B.K. Shoichet and I.D. Kuntz. "Protein docking and complementarity," *J. Mol. Biol.*, vol. 221, no. 1, pp. 327–346, 1991.

- [115] B.K. Shoichet. "Virtual screening of chemical libraries," *Nature*, vol. 432, pp. 862–865, 2004.
- [116] C.J.A. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, P. Bucher. "PROSITE: A documented database using patterns and profiles as motif descriptors," *Brief Bioinform.*, vol. 3, pp. 265–274, 2002.
- [117] C.V. Smith, C.C. Huang, A. Miczak, D.G. Russell, J.C. Sacchettini and K. Honer Zu Bentrup. "Biochemical and structural studies of malate synthase from *Mycobacterium tuberculosis*," *J.Biol.Chem.*, vol. 278, pp. 1735–1743, 2003.
- [118] V. Sobolev, R.C. Wade, G. Vriend, M. Edelman. "Molecular docking using surface complementarity," *Proteins Struct. Func. Genet.*, vol. 25, pp. 120–129, 1996.
- [119] I. Sommer, O. Muller, F.S. Domingues, O. Sander, J. Weickert and T. Lengauer. "Moment invariants as shape recognition technique for comparing protein binding sites," *Bioinformatics*, vol. 23, no. 23, pp. 3139–3146, 2007.
- [120] L. Song, C. Kalyanaraman, A.A. Fedorov, E.V. Fedorov, M.E. Glasner, S. Brown, H.J. Imker, P.C. Babbitt, S.C. Almo, M.P. Jacobson and J.A. Gerlt. "Prediction and assignment of function for a divergent N-succinyl amino acid racemase," *Nature Chemical Biology*, vol. 3, pp. 486–491, 2007.
- [121] M. Stahl and M. Rarey. "Detailed analysis of scoring functions for virtual screening," *J Med Chem.*, vol. 44, no. 7, pp. 1035–1042, 2001.
- [122] M. Stahl and H. Mauser. "Database clustering with a combination of fingerprint and maximum common substructure methods," *J. Chem. Inf. Model*, vol. 45, no. 3, pp. 542–548, 2005.

- [123] G.R. Stockwell, and J.M. Thornton. “Conformational diversity of ligands bound to proteins,” *Journal of Mol. Biol.*, vol. 356, pp. 928–944, 2006.
- [124] SYBYL 7.3, Tripos International, MO.
- [125] G. Szabo, J. Fischer, A. Kis-Varga and K. Gyires. “New celecoxib derivatives as anti-inflammatory agents,” *J Med Chem.*, vol. 51, no. 1, pp. 142–147, 2008.
- [126] C. Taroni, S. Jones and J.M. Thornton. “Analysis and prediction of carbohydrate binding sites,” *Protein Engineering*, vol. 13, no. 2, pp. 89–98, 2000.
- [127] J.W. Torrance, G.J Bartlett, C.T. Porter and J.M. Thornton. “Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families,” *J. Mol. Biol.*, vol. 347, no. 3, pp. 565–581, 2005.
- [128] R. Wang and S. Wang. “How does consensus scoring work for virtual library screening? An idealized computer experiment,” *J Chem Inf Comput Sci.*, vol. 41, no. 5, pp. 1422–1426, 2001.
- [129] S. Wang and D. Eisenberg. “Crystal structure of the pantothenate synthetase from *Mycobacterium tuberculosis*, snapshots of the enzyme in action,” *Biochemistry*, vol. 45, pp. 1554–1561, 2006.
- [130] Y. Wei, D. Ringe, M. Wilson, M. Ondrechen. “Identification of functional subclasses in the DJ-1 superfamily proteins,” *PLoS Computational Biology*, vol. 3, no. 1, 2007.
- [131] E.R. Zartler and J. Shapiro. “Fragonomics: Fragment-based drug discovery,” *Current Opinion in Chemical Biology*, vol. 9, pp. 366–370, 2005.
- [132] S. Zenno, H. Koike, A.N. Kumar, R. Jayaraman, M. Tanokura and K. Saigo. “Biochemical characterization of NfsA, the *Escherichia coli* major nitroreductase

- exhibiting a high amino acid sequence homology to Frp, a *Vibrio harveyi* flavin oxidoreductase,” *J Bacteriol.*, vol. 178, no. 15, pp. 4508–4514, 1996.
- [133] Y. Zheng, H. Li and D. Doermann. “Text identification in noisy document images using Markov random field,” in *Proc. Seventh International Conference on Document Analysis and Recognition*, vol. 1, pp. 599, 2003.
- [134] M. Zolli-Juran, J.D. Cechetto, R. Hartlen, D.M. Daigle and E.D. Brown. “High throughput screening identifies novel inhibitors of *Escherichia coli* dihydrofolate reductase that are competitive with dihydrofolate,” *Bioorganic & Medicinal Chemistry Letters*, vol. 13, pp. 2493–2496, 2003.
- [135] ” *Pubchem Database*,” <http://pubchem.ncbi.nlm.nih.gov/>

VITA

Reetal Pai Karkala earned her B.E. degree in electrical and electronic engineering from The National Institute of Engineering, Mysore, India. She obtained a M.S. degree from Clemson University in computer engineering under the guidance of Dr. Adam Hoover. She obtained her Ph.D. degree from Texas A&M University under the guidance of Dr. Thomas Ioerger. Her research interests are in the application of pattern recognition and machine learning techniques to solve problems in the field of biology and chemistry. She is currently working as a Senior Research Analyst at Dow Agrosiences where she is applying her research skills to the improvement of the corn germplasm.

Contact Address: Department of Computer Science and Engineering
Texas A&M University, TAMU 3112, College Station, TX 77843-3112