

USING ENERGY CENTRALITY RELATIONSHIP (ECR) TO  
IDENTIFY AND PREDICT FUNCTIONALLY-LINKED  
INTERACTING PROTEINS (FLIPS)

A DISSERTATION  
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY  
IN THE GRADUATE SCHOOL OF THE  
TEXAS WOMAN'S UNIVERSITY

DEPARTMENT OF BIOLOGY  
COLLEGE OF ARTS AND SCIENCES

BY  
SANJANA SUDARSHAN

DENTON, TEXAS

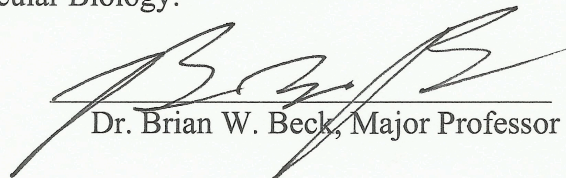
MAY 2015

TEXAS WOMAN'S UNIVERSITY  
DENTON, TEXAS

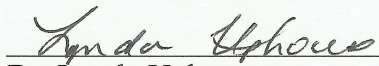
March 3, 2015

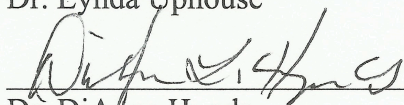
To the Dean of the Graduate School:


I am submitting herewith a dissertation written by Sanjana Sudarshan entitled "Using Energy Centrality Relationship (ECR) to identify and predict Functionally-Linked Interacting Proteins (FLIPs)". I have examined this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy with a major in Molecular Biology.

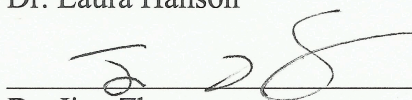
  
Dr. Brian W. Beck, Major Professor


We have read this dissertation and recommend its acceptance:

  
Dr. Lynda Uphouse

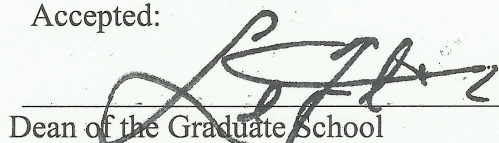
  
Dr. DiAnna Hynds

  
Dr. Laura Hanson

  
Dr. Jian Zhang

  
Department Chair

Accepted:

  
Dean of the Graduate School

## ACKNOWLEDGMENTS

I would like to take this opportunity to thank my advisor Dr. Brian W. Beck for his continuous support throughout my doctoral study. This work would not have been possible without his guidance and encouragement.

I would like to thank my committee, Drs. Hanson, Hynds, Uphouse, and Zhang for their valuable time and evaluation of my progress, during my time here.

I also thank my friends and lab mates, Amruta Mahadik, Sasi Kodathala, Isha Mehta, and Cameron Jones for their help at various stages of my graduate degree.

Finally, and most importantly, I would like to thank my family. Without the love and support from my family I would not have come this far.

## ABSTRACT

SANJANA SUDARSHAN

USING ENERGY CENTRALITY RELATIONSHIP (ECR)  
TO IDENTIFY AND PREDICT FUNCTIONALLY-LINKED  
INTERACTING PROTEINS (FLIPS)

MAY 2015

Interacting networks of proteins are responsible for a multitude of biological functions. These Functionally-Linked Interacting Proteins (FLIPs) occur at specific interfaces. It is therefore important to distinguish them from Functionally uncorrelated Contacts (FunCs). Here we utilize geometric, energetic, and sequence conservation characteristics at the interface to identify factors that may contribute towards an interface being FLIP or FunC. We studied these interface properties by analyzing a protein database we created called FLIPdb, which contains proteins belonging to various functional sub-categories. In our approach, which we term the Energy Centrality Relationship (ECR), we coupled Kortemme and Baker's computational alanine scanning analysis to estimate the energetic sensitivity of each amino acid at the center of the interface with geometric features. Principal Component Analysis and K-means Clustering analysis on FLIPdb could distinguish FLIPs from FunCs with an accuracy of 76%. To

investigate if evolutionary pressure plays a role in maintaining FLIPs, similar analyses were carried out on a set of 154 interfaces. Here we use Lichtarge's Evolutionary Trace (ET) method to calculate the ET score ( $\rho$ ) and alignment variability (# of states) of residues within various types of interfaces. Using PCA and K-means clustering analysis, we were able to distinguish FLIPs from FunCs with an accuracy of 69%. We also tested ECR's ability to identify near-native ( $\leq 5$  Å RMSD) poses in a docking run. A common problem in molecular docking is the generation of a large number of false positives. The ECR methodology was able to predict near-native poses in 50% of the cases, representing an increase of 9% relative to HEX (a well known docking software package) alone. Overall, we identified that FLIPs have a stronger central organizing tendency than FunCs. Although FLIPs also show more conservation at the core than at the edges, they exhibit more overall variability than FunCs, suggesting energy is conserved at the expense of sequence stability. Finally, we indicate how our ECR method may be used to reduce false positive predictions in docking calculations.

## TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS .....	iii
ABSTRACT .....	iv
LIST OF TABLES .....	x
LIST OF FIGURES .....	xi
 Chapter	
I. INTRODUCTION .....	1
Protein Interactions .....	1
Significance of Protein Interactions .....	2
Types of Protein Interactions .....	2
Homo- and Hetero-Oligomeric Complexes .....	2
Obligate and Non-Obligate Complexes .....	3
Transient and Permanent Complexes .....	3
Characterization of Protein-Protein Interfaces .....	5
Interface Centrality .....	6
Prediction .....	8
Overview of Findings .....	11
References .....	13
 II. PROTEIN-PROTEIN INTERFACE DETECTION USING THE ENERGY CENTRALITY RELATIONSHIP (ECR) CHARACTERISTIC OF PROTEINS .....	   18
Abstract .....	18
Introduction .....	19
Results .....	25
Database Composition, FLIPdb .....	25
CAS $\Delta\Delta G$ Distribution in PPI .....	26
Energy Centrality Hypothesis .....	28
Energetic And Geometric Features .....	28
Principal Component Analysis And K-Means Clustering .....	29

Accuracy And Matthews Correlation .....	31
Cross-validation Testing .....	31
Discussion .....	34
ECR Analysis Can Reproducibly Distinguish FLIP From FunC	
Interfaces .....	34
Physical Interpretation .....	35
Implications For Interface Evolution .....	36
Implications For Protein Docking .....	38
Conclusions .....	39
Methods .....	39
Dataset: FLIPdb .....	39
Computational Alanine Scanning (CAS) .....	44
Interfacial Geometry .....	44
Principal Component Analysis (PCA) .....	45
K-means Clustering .....	46
Accuracy And Matthews Correlation Coefficient .....	46
Acknowledgments .....	47
References .....	48

### III. FUNCTIONAL CLASSIFICATION OF PROTEIN INTERACTIONS USING INTERFACE SPATIAL DISTRIBUTION OF EVOLUTIONARY

CRITERIA .....	64
Abstract .....	64
Introduction .....	65
Results .....	70
Database Composition – FLIPdb-ET .....	70
$\rho$ Distribution In PPI .....	70
Evolutionary And Geometric features .....	72
Principal Component Analysis (PCA) .....	73
K-Means Clustering .....	73
Accuracy and Matthews Correlation .....	75
Cross-Validation Testing .....	75
Energetic and Evolutionary features .....	76
Discussion .....	78
Evolutionary Trace Can Distinguish FLIP From FunC Interfaces .....	78
Physical Interpretation .....	78
Implications .....	79
Conclusions .....	80
Methods .....	81
FLIPdb-ET Dataset .....	81
Evolutionary Trace .....	81
Interfacial Geometry .....	82
Geometry Analysis .....	82

Principal Component Analysis (PCA) .....	82
K-means Clustering .....	83
Accuracy and Matthews Correlation Coefficient .....	83
Acknowledgments.....	84
References.....	85
Table S3.1 References .....	98
IV. ENERGY CENTRALITY RELATIONSHIP REDUCES FALSE POSITIVE PREDICTION IN PROTEIN DOCKING.....	113
Abstract.....	113
Introduction.....	114
Results.....	117
ECR Prediction of Docking Targets .....	117
Representative Conformations.....	119
Easy Conformation: 1cdc .....	119
Intermediate Conformation: 1bsr .....	119
Hard Conformation: 1tub .....	120
Distribution of Poses.....	120
Symmetry Operator Prediction .....	121
Discussion .....	122
Docking Poses Exhibit Different Distribution Patterns .....	122
ECR Can Recognize Symmetry-Related Native Poses .....	122
HEX Struggles With Ab-Ag .....	125
Conclusion .....	125
Methods.....	126
Dataset: FLIPdb-lite.....	126
Molecular Docking .....	126
Computational Alanine Scanning (CAS).....	127
Interfacial Geometry .....	127
Energetic And Geometric Features.....	127
Principal Component Analysis (PCA).....	128
K-means Clustering .....	128
References .....	130
V. CONCLUSION.....	158
References .....	167
COMBINED REFERENCES .....	169

## APPENDIX

USING THE ENERGY CENTRALITY RELATIONSHIP IN A CAPRI-TYPE ANALYSIS OF UNBOUND TARGETS.....	185
--	-----

## LIST OF TABLES

Table	Page
2.1 Summary of Protein and Protein Interface Counts in FLIPdb .....	58
2.2 Accuracy of Clustering in Training and Test-18 Sets.....	59
S2.1 FLIPdb Interface Composition .....	61
S2.2 Summary of Protein and Protein Interface Counts in Dey-170 .....	62
S2.3 Pseudo-Accuracy of Clustering in Dey-170 Test Set.....	63
S2.4 Random Sub-Sample Validation of FLIPdb Training Set.....	64
3.1 FLIPdb-ET: Protein Interfaces and Functional Categories.....	89
3.2 Summary of Protein and Protein Interface Counts in FLIPdb-ET .....	90
3.3 Accuracy of Clustering in Training and Testing Sets .....	91
3.4 Accuracy of Clustering upon Combining ECR with ET .....	92
3.5 Random Sub-Sample Validation of FLIPdbET Training Set .....	93
S3.1 FLIPdb-ET Interface Composition.....	98
4.1 Docking Results of FLIPlite .....	135
4.2 Success of HEX Versus ECR .....	137
4.3 Analysis of Distribution Plots.....	138

## LIST OF FIGURES

Figure	Page
2.1. Distribution of Alanine Substitution Energies in FLIP and FunC Interfaces. ....	53
2.2. The Energy Centrality Relationship (ECR) for Interface Evolution .....	55
2.3. Correlation of Features with Principal Components.....	56
2.4. PCA and K-Means Clustering of Training and Test-18 Sets.....	57
S2.1. PCA and K-Means Clustering of The Dey-170 Set.....	60
3.1. Distribution of Conservation At Interfaces.....	94
3.2. Correlation of Features with Principal Components.....	95
3.3. PCA and K-Means Clustering of Training and Test Sets. ....	96
3.4. Correlation of Evolution and Energy Features with .....	
Principal Components .....	97
4.1. Distribution of Docking Poses of Representative PPIs in FLIPlite. ....	139
4.2. Histogram and Box Plots of RMSDs of Poses Relative to the Known Structure.....	140
4.3. ECR Prediction of 1tub .....	141
4.4. ECR Prediction of X1ubs1.....	142
S4.1. Distribution of Docking Poses of PPIs in FLIPlite.....	143
S4.2. Histogram and Box Plots of RMSDs of Poses Relative to the Known Structure. ....	154
A1.1. Docking Solution for Target 30.. .....	198
A1.2. ECR Analysis of the Docking Poses of Target 30. ....	199
A1.3. Docking Solution for Target 40. ....	200
A1.4. ECR Analysis of the Docking Poses of Target 40. ....	201
A1.5. Docking Solution for Target 41. ....	202
A1.6. ECR Analysis of the Docking Poses of Target 41. ....	203

## CHAPTER I

### INTRODUCTION

#### **Protein interactions**

Networks of interacting proteins carry out various biological functions. These interactions can be permanent, such as the proteins involved in the formation of molecular machinery, or transient, like many of the proteins involved in cell growth, signal transduction, enzyme and immune regulation, and cell adhesion<sup>1, 2</sup>. To understand how these molecular systems function, it is important to develop methods that identify and characterize protein interactions. The formation of protein interfaces is generally governed by electrostatics, hydrophobic interactions, shape complementarity (particularly Van der Waals [VdW] interactions), and the flexibility of the molecules involved<sup>3, 4</sup>. The chemical properties of the amino acid residues that comprise these interfaces and their interactions are therefore complex. Our understanding of these properties and their interplay is currently inadequate to use them as the basis for the prediction of which amino acid residues are important for the maintenance of protein structure as well as those that contribute to function<sup>5, 6</sup>. The research described here addresses this deficiency and proposes two related methods, our own *Energy Centrality Relationship* (ECR, described in Chapter 2) and the use of the *Evolutionary Trace* method (described in Chapter 3), to identify and classify such contacts. The efficacy of using these methods to augment other protein:protein interaction (PPI) prediction methods, particularly protein

docking, are explored in Chapter 4 and also serve to further support our concepts and methodology. The studies described here characterize the properties of interfaces (particularly at the amino acid level) and identify underlying geometric and energetic principles of amino acid residues across protein interfaces. It is our working hypothesis that identification of such features can improve the differentiation between functionally relevant and non-functional protein associations. In this work, we describe research that examines this general hypothesis.

### **Significance of protein interactions**

Through various biological and biochemical studies, it is evident that protein function is directly correlated to its quaternary structure. Understanding this structure-function relationship requires an in-depth structural characterization of protein interfaces. This would promote functional annotation of proteins of unknown function and help understand protein interactions at the system level. Overall, identifying interfacial features important to structure and function could lead to improved ways of manipulating signaling pathways by targeting interfaces with drugs designed for that purpose.

### **Types of protein interactions**

PPIs are very diverse, both structurally and functionally, and can be broadly classified into the following categories:

#### ***i. Homo- and hetero-oligomeric complexes***

PPIs that occur between identical chains are called homo-oligomers, while those occurring between non-identical chains are called hetero-oligomers<sup>7</sup>. Homo-

oligomers are generally easier to ascertain as they are permanent complexes that, in general, have large, hydrophobic surfaces and tend to bury a large portion of their hydrophobic residues<sup>5</sup>. Homo-oligomers are generally symmetrical, making them, in general, more stable complexes<sup>8</sup>. Many soluble and membrane proteins are seen to form homo-oligomers and are found in abundance in the Protein Data Bank<sup>9, 10</sup>. The structures of hetero-oligomers, on the other hand, are more difficult to determine as they can exist independently in solution and often form complexes according to the conditions of the solution<sup>7, 11</sup>.

**ii. *Obligate and non-obligate complexes***

When protomers (structural subunits) of a complex can exist independently *in vivo*, the complex is referred to as non-obligate. If the protomers are unstable on their own and not generally found in the cellular milieu, the complex is referred to as obligate. While functionally and structurally obligate interfaces generally form stable and permanent complexes, non-obligate interfaces can be transient or permanent<sup>12</sup>.

**iii. *Transient and permanent complexes***

Transient and permanent interactions are classified based on their stability and the lifetime of the complex. Obligate complexes usually form permanent interactions that are very stable and generally irreversible. In contrast, transient interactions can associate and dissociate *in vivo*. Transient interactions can further be classified into weak and strong interactions based on the affinity and stability of the interaction<sup>7, 13</sup>. A number of such interactions take place within the cell but not all lead to assemblies

that contribute to protein function. Functional protein assemblies require specific binding which is likely to be a result of selection pressures to maintain a functionally or structurally relevant interaction. Similarly, to avoid harmful aggregate formation, non-specific interfaces likely undergo negative selection<sup>14, 15</sup>. Better understanding of how selection pressures act on interfaces to maintain function may help us differentiate between functional and non-functional interfaces. Over evolutionary time, selective pressures could result in larger sequence conservation of residues in functional interfaces<sup>16</sup>.

The broad classification of protein interfaces discussed above often mixes structural and functional properties in their operational definitions. In an effort to separate these concepts, we have chosen to operationally define *Functionally-Linked Interfaces of Proteins* (FLIPs) (and the residues forming them) to be PPIs for which mutation or other chemical modification has been found to alter the native biological function. Similarly, we define PPIs that do not have such a known alteration in function as *Functionally uncorrelated Contacts* (FunC). There is a body of research, including our own work, that suggests no one particular physico-chemical characteristic occurring across the various different types of interfaces is sufficient to distinguish FLIPs from FunCs<sup>17-19</sup>. Such a lack of a defining characteristic likely arises because the combination of several traits tends to obscure the trends of any one individual feature that may be important<sup>17-19</sup>. In this work, we analyze a number these interface properties and correlate them to the propensity to differentially form FLIP and FunC interfaces. The physical and chemical

characteristics that are important in formation and maintenance of PPIs are discussed below.

### **Characterization of protein-protein interfaces**

The simplest definition of a PPI is distance based. Many approaches use an atomic distance cut-off criteria measured between the centers of atoms in each protein chain of a complex<sup>20-23</sup>. For example, an interface might be defined as those residues whose atoms are within 4.0 Å of the atoms of another protein chain. Another commonly used definition of the interface is based on Solvent Accessible Surface Area (SASA), which represents the amount of surface of a protein that is exposed to solvent (generally water)<sup>24, 25</sup>. Upon complexation, the total surface area of the individual monomers decreases and the difference in surface areas ( $\Delta$ ASA) represents the size of the interface.

Along with interface size, shape complementarity is a commonly used feature in interface prediction<sup>26</sup>. Interfaces are rarely tightly packed<sup>27</sup>, which is one of the reasons for greater stability at the interface. The stability of interfaces is also strongly influenced by hydrogen bond interactions, hydrophobic interactions, van der Waals forces, desolvation energy, and electrostatics at the interface. While electrostatics and desolvation primarily govern high affinity binding in proteins<sup>28</sup>, hydrogen bonds, van der Waal's forces and salt bridges often contribute significantly to the specificity of the many interactions<sup>29</sup>. While electrostatic interactions have a longer range than van der Waals forces and can therefore influence stability and affinity farther from an interface, they also tend to be much less numerous than van der Waals forces<sup>30</sup>.

## **Interface centrality**

Another feature that is useful in understanding interfaces is the identification of “core” (residues at the center of the interface) and “rim” (residues distal to the interface) regions. It has generally been seen that amino acids at the core are more hydrophobic than at the rim<sup>16</sup>. Core residues are frequently functionally and structurally important, and are also more conserved than rim residues<sup>31</sup>. Such functionally or structurally important residues are generally referred to as “hot spots”. Core residues also have differing contributions to binding and stability than rim residues and consequently have different evolutionary pressures for conservation<sup>16</sup>.

In Chapter 2, we explore the contribution of individual residues to interface binding stability. In particular, we characterize the sensitivity of interfaces to residue substitution with alanine. Though described in more detail in Chapter 2, our initial work indicated that residues nearer the Center of an Interface (CoI, the mean of all interfacial C $\alpha$  positions) tend to have different energetic contributions towards the maintenance of FLIP and FunC interfaces than do residues that are farther from the CoI. The analysis of preliminary data leads us to propose FLIPs generally exhibit a more radially symmetric (central) energy profile than FunCs. As some PPIs, particularly FLIPs, are the result of evolutionary pressure and selection, we also suggest there will be a detectable evolutionary “signature” in the form of distinguishable geometric and energy patterns at different interfaces. We speculate that the interactions start as weak contacts and “grow” (via advantageous mutations) in strength and size over evolutionary time. While amino

acid side chains contribute strongly to biological function, not all residues contribute equally<sup>32</sup>. We expect interfaces will show decreasing energetic and evolutionary contribution towards forming the interface as the distance from the CoI increases (See Figure 2.1). We define this as interface centrality. To study centrality, we took a two-pronged approach; i) correlate energy with distance from CoI, ii) correlate sequence conservation with distance from CoI.

We hypothesize that *analyzing the geometric centrality patterns of the physico-chemical properties of residues in PPIs will allow FLIPs and FunCs to be distinguished*. We further hypothesize that *proteins belonging to different functional sub-categories within the FLIP and FunC categories will cluster by function, even when the same general spatial and energetic features are used to characterize interactions*.

It is often challenging and laborious to obtain the quaternary structures that are required to address these hypotheses and study the molecular mechanisms of PPI formation. Most structures are obtained via X-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy<sup>33</sup>. It is therefore imperative to develop computational interface prediction methods that are fast and accurate. Protein docking is one such method that generates theoretical PPI structures (“poses”) using the structures of two or more monomers that are known to interact. An extremely small minority of the poses generated will reproduce known structures (true positives, both FLIP and FunC) while the vast majority will represent false positives<sup>34</sup>. We hypothesize that if the poses are false positives, they will not exhibit physical and biochemical properties like FLIPs and

are more likely to be similar to known FunCs. True positive poses with near native specific interactions would be expected to exhibit properties similar to known FLIPs. Thus, we expect that re-analyzing protein docking with respect to the putatively distinct physico-chemical properties of interface residues may lead to a reduction of the false positive rate and thus an overall improvement in the accuracy of docking methodologies.

### ***Prediction***

To distinguish between FLIP and FunC interfaces, various studies have analyzed different properties of the interface. Apart from the size of the interface, which is the chief discriminatory feature in many studies<sup>15, 35, 36</sup>, another commonly used feature is the amino acid composition. Only a fraction of residues, often referred to as hot spots, account for the majority of the binding energy and quaternary stability<sup>32</sup>. These hot spots can be identified experimentally by evaluating the free energy change by systematic amino acid mutation to alanine<sup>37</sup>. Loss of side chain functionality is assessed by a variety of biophysical methods<sup>38</sup>. The hot spots identified are sometimes deposited in databases such as ASEdb<sup>39</sup> and BID<sup>40</sup>; however, the complexity of protein interfaces and analysis of hundreds of variants makes identification of hot spots a laborious process and therefore rigorous computational methods are required<sup>41</sup>. Several groups have used energy-based features such as hydrogen bonds, solvation, and packing interactions<sup>23, 42-44</sup>, or non-covalent interactions<sup>45</sup>, to develop models to computationally detect hot spots.

Molecular dynamics (MD) simulations can describe changes in molecular conformations and particle interactions as a function of time and are commonly used to

characterize model systems<sup>46</sup>. The atomic level information gathered has been used to more accurately predict hot spots<sup>47</sup>. However, MD simulations are computationally expensive and cannot trivially be applied to large-scale predictions<sup>46</sup>.

While various groups have studied sequence conservation patterns of proteins, they have varying opinions of the relative conservation of interfaces and the rest of the protein<sup>45, 48-50</sup>. Ma and co-workers have shown a correlation between hot spots and conserved residues<sup>51</sup>. These hot spots are seen to form clusters and are either buried or bound to neighboring residues forming regions of hot residues with central residues being more conserved<sup>52, 53</sup>. Several groups have also shown that active site residues are conserved across protein families<sup>48, 54, 55</sup>. Despite various studies on residue conservation in proteins, it has been difficult to identify precise patterns of conservation features that allow for FLIP/FunC distinction. The use of sequence conservation alone for interface detection is still not sufficiently understood or developed that it can be used in the absence of other corroborative methods<sup>48, 50</sup>.

Another method commonly used to identify bound states of protein complexes is protein docking, a quaternary structure conformational sampling technique. Docking generates thousands of possible structures of receptor-ligand pairs by rotating and translating the conformations of an initial starting structure. Resulting conformational poses are then scored, generally based on energetics<sup>56</sup>, and ranked. However, docking has a two-part problem:

- i. Scoring functions are not fully capable of discriminating native structures

from ‘decoys’ (quaternary conformational poses far from the known native structure)<sup>57</sup>. As a result, docking generates a high number of false positives<sup>34</sup>. This may be due to the heavy reliance of docking algorithms on shape complementarity. While complementarity alone is sometimes sufficient to dock separated sub-units of a known complex, it is less successful at docking unbound complexes<sup>58</sup>.

- ii. Proteins are dynamic, and even though various groups have made attempts to account for flexibility in their docking algorithms, the effect of flexibility on protein docking needs to be studied further<sup>57, 59</sup>.

Since there has been partial but inconsistent success with the methods listed above, we approached this problem using a combination of specific variants of the methods. By using a hybrid method combining geometric relationships and various PPI features (e.g. CAS energy values) with multivariate statistical approaches, we sought to answer the following question:

*Can the physico-chemical properties of protein interactions discriminate FLIPs from FunCs?*

We approached this question with the following specific aims:

1. Identify patterns of physico-chemical properties of PPIs (particularly sensitivity to substitution and sequence conservation) that distinguish FLIPs from FunCs
  - a. Create and manually curate a set of known PPI structures available from the PDB into FLIP and FunC categories.

- b. Identify geometric, energetic, and evolutionary protein structural features that can discern FLIPs from FunCs.
2. Determine if the features identified in (1b) can also distinguish between different functional sub-categories within FLIPs
3. Identify FLIP-like structures from a set of docking decoys using this same methodology

### **Overview of findings**

We analyzed physico-chemical properties of protein interfaces including energy, sequence conservation, and geometric distribution of interface residues. We used Principal Component Analysis and K-means clustering analysis in a multi-stage approach towards interface distinction. To assess the usefulness of the methodology, docking analysis followed by post-filtering using our methodology was performed.

In Chapter 2, we describe the identification of energy related terms and their contribution to FLIP-FunC distinction in a dataset of 160 interfaces (FLIPdb).

Chapter 3 describes a similar study of sequence conservation related terms. There we analyze the role of evolutionary pressure on FLIP maintenance. We also study the combined contribution of interface energy and sequence conservation towards FLIP-FunC discrimination.

Chapter 4 describes docking analysis done on a subset of our FLIPdb interface dataset. We identify whether we can improve the accuracy of bound docking calculations by augmenting them with our concepts and methods.

Finally in Appendix 1, we performed a docking analysis similar to Chapter 4 but on interfaces originating from the CAPRI (Critical Assessment of Protein Interfaces) project<sup>60</sup>. Use of CAPRI targets allows us to assess the impact of conformational flexibility during PPI formation. We identify if we can improve the accuracy of “unbound” docking calculations by augmenting them with our method.

## References

1. Phizicky EM, Fields S (1995) Protein-protein interactions: methods for detection and analysis. *Microbiol Rev* 59:94-123.
2. Massova I, Kollman PA (1999) Computational Alanine Scanning To Probe Protein-Protein Interactions: A Novel Approach To Evaluate Binding Free Energies. *J Am Chem Soc* 121:8133-8143.
3. Bahadur RP, Zacharias M (2008) The interface of protein-protein complexes: analysis of contacts and prediction of interactions. *Cell Mol Life Sci* 65:1059-1072.
4. Valdar WS, Thornton JM (2001) Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* 42:108-124.
5. Jones S, Thornton JM (1996) Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* 93:13-20.
6. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. *J Mol Biol* 285:2177-2198.
7. Nooren IM, Thornton JM (2003) Diversity of protein-protein interactions. *EMBO J* 22:3486-3492.
8. Andre I, Strauss C, Kaplan D, Bradley P, Baker D (2008) Emergence of symmetry in homooligomeric biological assemblies. *Proceedings of the National Academy of Sciences of the United States of America* JID - 7505876.
9. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235-242.
10. Duran AM, Meiler J (2013) Inverted Topologies in Membrane Proteins: A Mini-Review. *Comput Struct Biotechnol J* 8:e201308004-e201308004.
11. Ezkurdia I, Bartoli L, Fariselli P, Casadio R, Valencia A, Tress ML (2009) Progress and challenges in predicting protein-protein interaction sites. *Brief Bioinform* 10:233-246.
12. Amoutzias G, Van de Peer Y Single-Gene and Whole-Genome Duplications and the Evolution of Protein-Protein Interaction Networks In: Anonymous (2010) *Evolutionary Genomics and Systems Biology*, John Wiley & Sons, Inc., pp 413-429.

13. Nyfeler B, Michnick SW, Hauri H (2005) Capturing protein interactions in the secretory pathway of living cells. *Proc Natl Acad Sci U S A* 102:6350-6355.
14. Madaoui H, Guerois R (2008) Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proceedings of the National Academy of Sciences* 105:7708-7713.
15. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2004) A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol* 336:943-955.
16. Guharoy M, Chakrabarti P (2005) Conservation and relative importance of residues across protein-protein interfaces. *Proc Natl Acad Sci U S A* 102:15447-15452.
17. Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. *J Mol Biol* 280:1-9.
18. Keskin O, Gursoy A, Ma B, Nussinov R (2008) Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chem Rev* 108:1225-1244.
19. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. *J Mol Biol* 285:2177-2198.
20. Ofra Y, Rost B (2003) Analysing six types of protein-protein interfaces. *J Mol Biol* 325:377-387.
21. Larsen TA, Olson AJ, Goodsell DS (1998) Morphology of protein-protein interfaces. *Structure* 6:421-427.
22. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R (1996) A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J Mol Biol* 260:604-620.
23. Kortemme T, Baker D (2002) A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A* 99:14116-14121.
24. Zhu H, Domingues FS, Sommer I, Lengauer T (2006) NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics* 7:27-27.
25. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2003) Dissecting subunit interfaces in homodimeric proteins. *Proteins* 53:708-719.

26. Laskowski RA (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* 13:323.
27. Hubbard SJ, Argos P (1994) Cavities and packing at protein interfaces. *Protein Sci* 3:2194-2206.
28. Zacharias M (2005) ATTRACT: protein-protein docking in CAPRI using a reduced protein model. *Proteins* 60:252-256.
29. Honig B, Yang AS (1995) Free energy balance in protein folding. *Adv Protein Chem* 46:27-58.
30. Fernández-Recio J, Totrov M, Abagyan R (2003) ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins* 52:113-117.
31. Levy ED (2007) PiQSi: protein quaternary structure investigation. *Structure* 15:1364-1367.
32. Wells CT (1995) A hot spot of binding energy in a hormone-receptor interface. *Science* 267:383-386.
33. Li B, Kihara D (2012) Protein docking prediction using predicted protein-protein interface. *BMC Bioinformatics* 13:7-7.
34. Shoichet BK, McGovern SL, Wei B, Irwin JJ (2002) Lead discovery using molecular docking. *Curr Opin Chem Biol* 6:439-446.
35. Henrick KE (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 372:774-797.
36. Elcock AH, McCammon JA (2001) Identification of protein oligomerization states by analysis of interface conservation. *Proc Natl Acad Sci U S A* 98:2990-2994.
37. Weiss MS, Metzner HJ, Hilgenfeld R (1998) Two non-proline cis peptide bonds may be important for factor XIII function. *FEBS Lett* 423:291-296.
38. Morrison KL, Weiss GA (2001) Combinatorial alanine-scanning. *Curr Opin Chem Biol* 5:302-307.
39. Thorn KS, Bogan AA (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* 17:284-285.

40. Fischer TB, Arunachalam KV, Bailey D, Mangual V, Bakhru S, Russo R, Huang D, Paczkowski M, Lalchandani V, Ramachandra C, et al (2003) The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics* 19:1453-1454.
41. DeLano WL (2002) Unraveling hot spots in binding interfaces: progress and challenges. *Curr Opin Struct Biol* 12:14-20.
42. Kim KT (2004) Computational alanine scanning of protein-protein interfaces. *Sci STKE* 2004:pl2-pl2.
43. <http://rosetta.bakerlab.org>.
44. Tuncbag N, Kar G, Keskin O, Gursoy A, Nussinov R (2009) A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief Bioinform* 10:217-232.
45. Choi YS, Yang J, Choi Y, Ryu SH, Kim S (2009) Evolutionary conservation in multiple faces of protein interaction. *Proteins* 77:14-25.
46. Adcock SA, McCammon JA (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem Rev* 106:1589-1615.
47. González-Ruiz D, Gohlke H (2006) Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding. *Curr Med Chem* 13:2607-2625.
48. Grishin NV, Phillips MA (1994) The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Sci* 3:2455-2458.
49. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci* 13:190-202.
50. Valdar WS, Thornton JM (2001) Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* 42:108-124.
51. Ma B, Elkayam T, Wolfson H, Nussinov R (2003) Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A* 100:5772-5777.

52. Keskin O, Ma B, Nussinov R (2005) Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol* 345:1281-1294.
53. del Sol A, O'Meara P (2005) Small-world network approach to identify key residues in protein-protein interaction. *Proteins* 58:672-682.
54. Ouzounis CF, Perez-Irratxeta CF, Sander CF, Valencia A (1998) Are binding residues conserved? Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing JID - 9711271.
55. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM (2002) Analysis of catalytic residues in enzyme active sites. *J Mol Biol* 324:105-121.
56. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics* 47:409-443.
57. Aytuna AS, Gursoy A, Keskin O (2005) Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics* 21:2850-2855.
58. Smith GR, Sternberg MJ (2002) Prediction of protein-protein interactions by docking methods. *Curr Opin Struct Biol* 12:28-35.
59. Erickson JA, Jalaie M, Robertson DH, Lewis RA, Vieth M (2004) Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J Med Chem* 47:45-55.
60. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJE, Vajda S, Vakser I, Wodak SJ (2003) CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* 52:2-9.

CHAPTER II

PROTEIN-PROTEIN INTERFACE DETECTION USING THE  
ENERGY CENTRALITY RELATIONSHIP (ECR)  
CHARACTERISTIC OF PROTEINS

A paper published in PLOS One (2014)

Sanjana Sudarshan<sup>1</sup>, Sasi B. Kodathala<sup>1</sup>, Amruta C. Mahadik<sup>1</sup>, Isha Mehta<sup>1</sup>,  
Brian W. Beck<sup>1,2,3†</sup>

**Abstract**

Specific protein interactions are responsible for most biological functions. Distinguishing Functionally Linked Interfaces of Proteins (FLIPs), from Functionally uncorrelated Contacts (FunCs), is therefore important to characterizing these interactions. To achieve this goal, we have created a database of protein structures called FLIPdb, containing proteins belonging to various functional sub-categories. Here, we use geometric features coupled with Kortemme and Baker's computational alanine scanning method to calculate the energetic sensitivity of each amino acid at the interface to substitution, identify hot spots, and identify other factors that may contribute towards an interface being FLIP or FunC. Using Principal Component Analysis and K-means clustering on a training set of 160 interfaces, we could distinguish FLIPs from FunCs

with an accuracy of 76%. When these methods were applied to two test sets of 18 and 170 interfaces, we achieved similar accuracies of 78% and 80%. We have identified that FLIP interfaces have a stronger central organizing tendency than FunCs, due, we suggest, to greater specificity. We also observe that certain functional sub-categories, such as enzymes, antibody-heavy-light, antibody-antigen, and enzyme-inhibitors form distinct sub-clusters. The antibody-antigen and enzyme-inhibitors interfaces have patterns of physical characteristics similar to those of FunCs, which is in agreement with the fact that the selection pressures of these interfaces is differently evolutionarily driven. As such, our ECR model also successfully describes the impact of evolution and natural selection on protein-protein interfaces. Finally, we indicate how our ECR method may be of use in reducing the false positive rate of docking calculations.

## **Introduction**

Proteins interact with and bind to other proteins forming both transient and long-term networks of specific complexes whose interfaces have highly-specific amino acid interactions [1-6]. These interfaces play vital roles in biological functions such as signal transduction, enzyme and immune regulation, adhesion, force generation, and maintenance of cellular structure. Methods for the identification and characterization of protein-protein interactions (PPIs) are thus critical to understanding how living systems function.

Development of experimental and computational techniques to identify PPIs has

shed light on the determinants of specific interactions, as well as on some general features for different types of interactions [2-5, 7-13]. Experimental high throughput screening methods [3-5, 14] have provided information to construct large databases [15-17] of PPIs and related functions. Computational methods such as molecular modeling and docking, have generally identified the shape, electrostatic complementarity, buried surface area, flexibility, solvation energy, and sequence conservation of the interactors (amino acid residues) as key features in interface detection [6, 7, 11-13, 18-20]. Use of these known relationships to better elucidate the principles by which amino acids are positionally organized and thus contribute energetically to interfaces would allow specific structure/function relationships to be characterized. Such knowledge could also promote the finding of novel interfaces via computational docking calculations, as well as allowing the testing of rival protein structure/function hypotheses. Unfortunately, the different attempts at characterization continue to be hampered by a fundamental lack of understanding about the underlying geometric and energetic principles of amino acid interaction across protein interfaces [6, 8, 19, 21-24].

Several potential reasons for this exist. Both experimentally and computationally, it has been observed that few of the residues present in a PPI are essential for maintenance of the integrity of the interface [2, 8, 24]. Some success has been had identifying these important hot spots, particularly with computational alanine scanning methods (CAS) [2, 25-29]. However, the use of CAS in PPI detection has had mixed success. CAS methods often very accurately distinguish residues critical to known

interfaces, while failing to identify all the residues in an interface [24]. Ofra and colleagues suggest that this may be due, in part, to a bias towards hot spot residues that may treat non-hot spot residues as “noise” and thus fail to identify all the residues in a PPI [24].

An additional reason PPI principles may be difficult to elucidate can be found in how the experimental data used to develop computational methods like docking is organized and utilized. Most data for the patterns of amino acid characteristics at PPIs come from atomic resolution structures of protein complexes deposited at the Protein Data Bank (PDB) [30]. While an understanding of PPI principles for both prediction and design necessitates the use of natural exemplars, whether a reference structure is a highly specific interaction used in nature and critical for a biological function or whether the association is the result of the experimental conditions used in the technique can often be unclear. The majority (approximately 80%) of PPI structures available from the PDB are obtained through X-ray crystallography [31]. The very symmetrical and tightly packed structures that promote facile structure determination can also indicate interfaces not present in the cellular milieu [23, 32, 33]. As with hot spot/non-hot spot bias, development of PPI predictive methods based simultaneously on both aggregative (e.g. crystal contacts) and functionally-linked PPIs may obscure trends such that both can fail to be identified.

Several groups have classified PPIs into different operationally defined categories

such as, homo- and hetero- complexes, obligate and non-obligate complexes, and transient and permanent complexes (reviewed in [6, 34]). These categories, however, often mix structural and functional properties in their operational definitions. While structure and function are, of course, related, natural selection operates on biological function, and it may serve useful to identify the functional importance of a given PPI as a separate characteristic feature. In this work, we operationally define Functionally-Linked Interfaces of Proteins (FLIP), and the residues forming these interfaces, to be those for which mutation or other chemical modification has been found to alter the native biological function. Similarly, we define PPIs that do not have such a known alteration in function as Functionally uncorrelated Contacts (FunC).

Separation of FLIPs from FunCs can be problematic using PDB data alone, and additional knowledge is generally required [7, 13, 35] FLIPs and FunCs can be thought of as positive-design (specific) and negative-design (aggregative) natural exemplars in the parlance of Havranek [36]. While the PDB often provides a “Biological Assembly” structure (BioUnit) in addition to the standard “Asymmetric Unit” structure, in our experience, the correlation of the BioUnit structures with FLIPs is not straightforward. BioUnits are often not available, are duplicates of the Asymmetric Unit with little justification for that assignment, or are specified for non-native interactions as in the case of rabbit actin with bovine DNase (PDBid: 1ATN). As mentioned previously, shape and electrostatic complementarity, buried surface area, flexibility, solvation energy, amino acid composition, hydrophobicity, and sequence conservation are all common features

used to characterize and predict the quaternary assemblies and improve estimation of likely solution state structures [7, 11-13, 20, 37]. Indeed, more recent BioUnit assignments have been improved through the automated use of tools like PISA, which has a particular strength in that it leverages solvation energy calculations in addition to other features to identify macromolecular complexes in solution [13]. Even with these enhanced analyses, the relationship of the complex with function may still be problematic. For example, PISA, NOXclass, and EPPIC servers all identify Actin:DNase as the likely BioUnit [11-13]. As a result, the ability to distinguish FLIP from FunC, though improved, remains obscure. While large interactome databases exist that often do indicate functional correlation [15-17], they generally specify whole protein chain or complex interactions and do not specify data at the atomic level.

In principle, it is possible to use atomistic or coarse-grain computational methods, including docking methods, that use generic, empirical amino acid interaction functions to successfully predict quaternary interactions [19, 29, 38-40]. Unfortunately, two problems generally arise: 1) the false positive rate (average number of predictions needed to obtain a structure similar to a natural exemplar) is fairly high [38, 41, 42] and 2) while accurate structures can be identified, assessment as to the functional significance (i.e. FLIP or FunC) is not generally identified or remains obscure [19, 23, 38, 42].

Physico-chemical properties of the amino acid residues in PPIs other than sensitivity to alanine substitution have also been investigated, including hydrophobicity, amino acid composition, hydrogen bonding potential, sequence conservation, and solvent

accessible surface area (SASA), all with differing success [6, 23, 42]. Combining these methods in hybrid approaches has improved successful identification of native PPIs relative to any one property alone [6, 11, 13, 23].

In light of these improvements, a hybrid approach that includes the statistical analysis of (a) atomic-resolution interface geometries and (b) CAS-based energy data of protein structures pre-classified based on functional importance (FLIP/FunC) may be successful, both in improving detection of interfaces and increasing our understanding of general principles of interface formation. To test this concept, we collected a set of PPI structures available in the PDB starting from a subset of members of commonly used sets to test PPI and docking software [7, 20, 43-45] and added additional structures of interest to the lab. We then used additional literature sources to manually categorize the interfaces as being FLIP or FunC (FLIPdb, see Methods). For each interface in FLIPdb, we used Baker's CAS method [28] and our own geometry calculations (see Methods) to determine the energetics of alanine substitution of residues in a PPI as a function of geometric distribution in the interface. No attempt was made to bias towards only hot spot data. Using Principal Component Analysis [46] and K-means clustering [47] we were able to identify seven physical characteristics that could distinguish FLIP interfaces from FunC interfaces with 76% accuracy. These same characteristics, when tested against a set of 18 unrelated PPI structures and a subset of 170 PPI from the set of Dey et al., were also able to distinguish FLIP from FunC with 78-80% accuracies. Overall, FLIP interfaces appear to have greater overall sensitivity to ala substitution than FunC (Figures 2.1-2.4),

*particularly toward the center of the interfaces*. This may be related to the finding that cores of interfaces have greater sequence conservation than interfaces rims [48]. Both are consistent with the ideas that FLIP interfaces are more specific than FunC interfaces [1, 6] and that they may evolve increasing specificity radially across a PPI over evolutionary time (Figure 2.2a-c).

The novelty of this approach, which we term the Energy Centrality Relationship (ECR), is that through the combination of geometric and energetic data, we are able to not only reproduce functional classifications and describe physical chemical sources of these differences, but also have a model that is consistent with natural selection pressures on protein interfaces.

## **Results**

### Database Composition, FLIPdb

After construction, our FLIPdb database referenced 160 PPIs between 233 protein chains that were contained in 94 PDB structural files. This set was categorized and divided into 100 FLIP interfaces and 60 FunC interfaces. We further sub-categorized these PPIs into 7 FLIP and 2 FunC sub-categories: 1) antibody-antigen (AbAg); 2) immunoglobulin Heavy Chain/Light Chain (AbHL); 3) Enzyme-Enzyme, both transient and persistent (Enzyme); 4) having a generally persistent structure that provides mechanical stability, such as cytoskeletal or viral proteins (Structural); 5) peptide/protein inhibitors to an enzyme (Inhibitor); 6) proteins whose function is to recognize

peptides/proteins (Receptor); 7) proteins regulated by post-translational modification by another protein (Regulated); 8) PPIs in an asymmetric crystal unit NOT found to be FLIP (FunC); and 9) PPIs obtained by applying crystal symmetry transforms to FLIP structures (XFunC). This set of PPIs (see supplementary Table S2.1) was used for training and development (summary in Table 2.1).

An additional set of 18 PPIs between 19 protein chains in 7 PDB files was also categorized into 13 FLIP and 5 FunC interfaces and sub-categorized as above (see supplementary Table S2.1). This second set of PPIs was comprised of proteins that were generally less than 70% sequence identical to proteins in the training set and was used for cross-validation testing (Table 2.2).

Finally, a third set of 170 PPIs between 301 protein chains in 139 PDB files was examined. These 170 PPIs represent a subset of 54% of the weakly and strongly interacting PPIs characterized by Dey and colleagues [7]. This set was not rigorously curated as to FLIP/FUNC status so as to compare the results of our training set with that of Dey. Overall, the structures and energetics of 348 PPIs were categorized and examined.

#### CAS $\Delta\Delta G$ Distribution in PPI

We used Baker's CAS method [28] coupled with our own software to determine the sensitivity to alanine substitution of residues in a PPI, as a function of geometric distribution in the interface. All geometric analyses were based on residue C $\alpha$  positions.

This sensitivity was compared between FLIP and FunC PPIs in the FLIPdb. Two representatives of this are shown in Figure 2.1, in which we compare a FunC (Yeast Phosphotransferase Ypd1p, PDBid: 1C02) and a FLIP (T4 bacteriophage dC-hydroxymethylase dimer, PDBid: 1b5e). Histogrammed contours of the pseudo-free energy change upon alanine substitution ( $\Delta\Delta G$ ) are plotted on the principal component analysis (PCA) projections of the interface residue geometry (Figures 2.1a,b). (Note, that in this work, we follow Baker's use of the terms "free energy" and " $\Delta G$ " for consistency with the software output.) These distributions indicate that in the FLIP, "hotter" residues (whose CAS analysis resulted in more PPI destabilization upon substitution) tend to be more centrally located and tend to show a progressive radial symmetry. In contrast, the "hotter" residues in the FunC are fairly evenly distributed throughout the interface. Some "cold" residues (those favoring Ala substitution) are found near the interface center. These CAS energy distributions are representative of other FLIPs and FunCs. When all the  $\Delta\Delta G$  vs. distance from the Center of Interface (CoI) were then fitted to a 1st order polynomial line via linear regression, 8 of the 10 highest intercepts were found to be FLIP, while 8 of 10 lowest intercepts were found to be FunC (Figure 2.1c). In general, FLIPs were found to fit a line better than the FunC (coefficients of determination,  $R^2$ , were an order of magnitude larger). The FLIPs were also found to generally have a negative slope, indicative of a central tendency, whereas the FunCs generally had near flat or small magnitude positive slopes. The small magnitude slope and poor  $R^2$  suggests little geometric central tendencies in the FunC. These trends were generally maintained

throughout FLIPdb, with most FLIPs having a radially symmetrical central tendency and most FunCs demonstrating little-to-no correlation with distance from the center of the interface. One-way pairwise ANOVA at an  $\alpha = 0.10$  analyzing the slopes and intercepts indicated that the differences between FLIP and FunC were significant with  $P \leq 0.0006$  and  $P \leq 0.09$ , respectively.

### Energy Centrality Hypothesis

There is no *a priori* reason FLIP PPIs should demonstrate a central tendency relative to FunC PPIs. Unless an organizing principle was involved, one might expect an interface to have a random correlation between CAS  $\Delta\Delta G$  and geometry (Figure 2.2c-d). The presence of such a central tendency (Figure 2.1) in FLIP interfaces suggests that they are indeed organized (Figure 2.2e-f), perhaps through a natural selection process (see Discussion and Figure 2.2).

### Energetic and Geometric Features

Though PPIs are complex 3-dimensional entities, for the sake of simplicity of analysis, we unified CAS  $\Delta\Delta G$  and structural geometry characteristics into scalar quantities that could be used to describe a PPI. Three features arose from the regression of energy to geometry: the rate of change of substitution energy as a function of distance ( $\Delta r$ ) from the interface center (slope\_ $\Delta\Delta G$ ), the extrapolated maximum  $\Delta\Delta G$  sensitivity at the interface center (intcpt\_ $\Delta\Delta G$ ), and the adherence of the  $\Delta\Delta G$  and  $\Delta r$  data to a linear relationship (coefficient of determination, R<sup>2</sup>\_ $\Delta\Delta G$ ). Three features were found that

describe the net sensitivity of an interface to CAS: net sum of all  $\Delta\Delta G$  changes (Sum\_ $\Delta\Delta G$ ), mean  $\Delta\Delta G$  for all interface residues (Avg\_ $\Delta\Delta G$ ), and total number of residues in the interface (#total). The remaining two features address the number of residues extremely sensitive to Ala substitution (“hot” residues, residues with  $\Delta\Delta G$  larger than +1 kcal/mol): the number of hot residues (#hot), and the ratio of hot to total (frac\_hot). One-way pairwise ANOVA at an  $\alpha = 0.10$  indicated that all features except R2\_ $\Delta\Delta G$  were significantly different between FLIP and FunC with #hot, total, Sum\_ $\Delta\Delta G$ , frac\_hot, and Avg\_ $\Delta\Delta G$  having  $P \leq 0.0001$ , intcpt\_ $\Delta\Delta G$  having  $P \leq 0.0006$ , and slope\_ $\Delta\Delta G$  having  $P \leq 0.09$ . Since these features could reasonably be viewed as coupled, we also performed one-way ANOVA with repeated measure at an  $\alpha = 0.10$  and with Tukey-Kramer post-hoc analysis. This analysis indicated differences between FLIP and FunC for #hot, total, Sum\_ $\Delta\Delta G$  that were significantly different with  $P \leq 0.0001$ . Though shown to be statistically different, individually none of these features were found to sufficiently correlate with FLIP or FunC categories such that a single feature could be used to identify the category.

### Principal Component Analysis and K-Means Clustering

When no single feature could easily discriminate FLIP from FunC, yet each feature yielded significant differences between groups, the multi-factoral approach of PCA was used. Initial PCA analysis of the 8 features for all 160 PPI in the training set yielded a set of principal components (PCs) that reproduced 80% of the normalized data variation in the first two PCs (Figure 2.3a). Analysis of the eigenvector coefficients

(Figure 2.3a) agreed with the ANOVAs indicating that the variance in the data was far less dependent on a strict adherence to a 1<sup>st</sup> order linear model. Thus, for all subsequent analyses, R2\_ΔΔG was dropped as a feature. The resultant 7-feature PCA reproduced 88% of the remaining data variation in the first two PCs (Figure 2.3b). Subsequent K-means cluster analysis with a two-cluster assumption of this data (Figure 2.4a) produced two clusters whose centroids straddled the origin for both PC1 and PC2 indicating opposing correlation trends. Analysis of these clusters revealed they had high *precision* and *specificity*. Cluster 1 contained 49% of all FLIPs but only 2% of FunC PPI (Table 2.2). Cluster 2 contained 51% of all FLIPs and 98% of FunC PPI. The FLIP PPI in cluster 1 were predominately in the Enzyme (72% of Enzyme) and Antibody-Heavy/Light sub-categories (100% of AbHL). Cluster 2 was dominated by FunC/XFunC (98% of FUNC), Antibody-Antigen (75% of AbAg), and Inhibitor sub-categories (100% of Inhibitor). Closer examination of cluster 2 revealed that FLIPs assigned to this cluster tended towards more positive PC1 values and larger magnitude PC2 values than FunCs/XFunCs. This consistency in trend suggested a second PCA over the same features might provide further distinction between FLIPs and FUNCs. A new PCA of only the 110 PPI in cluster 2 of the first PCA produced new PCs with extremely similar eigenvector coefficient correlations to the first PCA (Figure 2.3c). The same set of features still produced PCs that represented 84% of the resultant data variation in the first two PCs. This confirmed that similar data dependencies were in effect between the two PCA. K-means clustering of this second PCA again produced 2 clusters that straddled the

origin for both PC1 and PC2 (Figure 2.4b). As with the first PCA, cluster 1 of the second PCA was predominately FLIP, containing 61% of the remaining FLIPs but only 28% of the total FunCs. Likewise, cluster 2 was predominately FunCs, containing 20% of the FLIPs and 70% of the FunCs PPI (Table 2.2). Over two-rounds of PCA, 80% of the FLIPs were found in the clusters positively correlated with the features, and 70% of the FunCs were found in clusters negatively correlated with the features.

#### Accuracy and Matthews Correlation

Analysis of the two rounds of PCA of the training set PPI data indicated that the overall accuracy (the propensity to correctly identify FLIP or FunC) was ~67% in each PCA round. Over both rounds of PCA, the accuracy was 76% (Table 2.2). The Matthews Correlation Coefficient, a measure of how well a binary classification matches the data, was 0.49 in PCA round one, 0.32 in PCA round two, and 0.50 across both rounds. Such MCCs indicate a two-category assumption is quite consistent with the data.

#### Cross-validation Testing

While analysis of the training set data very favorably predicted distinct feature set correlations between FLIPs and FunCs, it was possible that the relationship was training set dependent and demonstrated compositional bias. In order to test this, we undertook three types of cross-validation testing: validation on two test sets and random sub-sampling validation on the training set.

We first repeated the analyses on the 18 member test set (hereafter, Test-18). The additional interfaces in this set were between protein chains that generally had less than 70% sequence identity to chains in the training set (Table S2.1, see Methods). No new PCA or K-means clustering was undertaken; rather the features of Test-18 were projected through the PCs of the training set. Test-18 projections are shown in Figures 2.4c,d. As with the training set, FLIPs in the Enzyme and Antibody-Heavy/Light sub-categories could be reliably identified in cluster 1 of PCA round 1. Similarly, FunCs dominated the composition of cluster 2 in PCA round 2. While the accuracies of the PCA 1 projection were disappointingly lower than the training set (48%), the 2<sup>nd</sup> round projection accuracies were larger (73%), and the overall two-round accuracy was actually slightly higher than the training set at 78% (Table 2.2). Similarly, MCC values were also slightly higher, at 0.62 (Table 2.2). This backhanded success may in part arise due to the relatively high fraction of AbAg in Test18, as AbAg are generally identified in round 2.

We next repeated the analyses on a second test set of 170 PPI derived from the dataset of Dey and colleagues (see Methods) [7]. The dataset of Dey and colleagues was designed to analyze PPI known to interact weakly or strongly in solution. Our subset (hereafter Dey-170) represents about 54% of the full Dey dataset and contains 32 weakly interacting PPI (weak) and 138 strongly interacting PPI (strong) (Table S2.2). Dey-170 was not rigorously curated as to FLIP/FUNC status but instead was used to examine two model assumptions: a) Assume all 170 PPI are FLIP-like since all are known to oligomerize in solution or b) Assume weak PPI are more FUNC-like and strong PPI are

more FLIP-like. Testing these assumptions allows us to examine how well our operationally defined categories of FLIP and FUNC agree with the weak and strong PPI characterized by Dey. Again, the values of the 7 features of each Dey-170 PPI were projected through the PCs of the training set (Figure S2.1a,b, Table S2.3). In projection round 1, cluster 1 contained 59% of the strong PPI and no weak PPI. Cluster 2 contained 100% of the weak PPI and 41% of the strong PPI. In round 2, 75% of the remaining strong PPI and 38% of the weak PPI were found in cluster 1, while 62% of weak PPI and 10% of strong PPI were found in cluster 2. If we follow crude assumption (a) that all Dey-170 are FLIP (i.e. no true negatives or false positives exist), we still achieve an overall accuracy of 80% (Table S2.3a). As this assumption is false, this accuracy likely represents a lower limit. Interestingly, though this assumption has a near zero MCC (random guessing) in round 1, subsequent rounds of projection positively improve the correlation to an overall MCC of 0.12. The accuracy and improving MCC suggest that a two-category model, even when mis-assigned is superior to random chance. If we follow crude assumption (b) that weak PPI are FUNC-like and strong PPI are FLIP-like, we obtain results consistent and slightly superior to the training set results with accuracies of 84.7% and an MCC of 0.51 (Table S2.3b).

As the accuracy and MCC varied somewhat from training set to Test-18 set to Dey-170 set, we evaluated the compositional bias of our training set using random sub-sampling validation (Table S2.4). Sub-samples of the training set were generated randomly in triplicate for subsets of the training set ranging from 90% down to 20%.

Regression analysis at an  $\alpha = 0.10$  for 1<sup>st</sup> through 6<sup>th</sup> order polynomial fits of number of PPI vs. Accuracy show substantial *Lack of Fit* error and a lack of statistical significance for each. Overall, while this suggests that little compositional bias exists until the number of PPI falls substantially below 80 (50% of the training set), it also suggests that analyzing more PPI will not dramatically improve the overall accuracy.

Taken together these training set and random sub-sampling results suggest our method is robust to protein identity and of general applicability, though likely needing additional refinement in order to boost the accuracy to levels found in other methods [11-13]

## **Discussion**

### **ECR Analysis can Reproducibly Distinguish FLIP from FunC Interfaces**

Through the coupling of biological functional categorization with interface geometries and energetics, the ECR methodology produces very consistent results, both between training and testing sets, as well as between functional sub-categories of PPI. FLIP PPIs can be distinguished from FunC PPIs with 76% accuracy (Table 2.2). In addition, PPIs of the same functional sub-category generally have similar PC projection values such that they cluster (Figures 2.4 & S2.1). An accuracy of 76% compares favorably with other approaches combining several methods [19, 23, 24]. It has slightly lower accuracies (by approximately 10-12%) than PISA, NOXclass, and EPPIC [11-13]. While lower in overall accuracy than some of the most accurate methods, it does not

appear to have any significant compositional bias. ECR also has a distinct advantage over many methods in that it is based solely on interaction energies and structural features and does not rely on sequence conservation patterns or interactome maps [15-17]. However, given the success of approaches like those above that use sequence conservation, particularly sequence entropy, we can expect that future inclusion of features from these other approaches in our analysis would not hinder and might even improve our accuracy. Furthermore, the reproducibility across functional sub-categories, a characteristic not included in the model but rather emergent from the analysis, suggests that this method may also be useful in the annotation of PPIs with unknown function. It is also an improvement on methods that rely solely on hot spot analysis in that through examination of all interface residue interactions, ECR provides an energetic context for the hot spots and their differential presence in FLIP and FunC PPIs.

### Physical Interpretation

From the analysis of CAS  $\Delta\Delta G$  energetic and geometric features, several clear patterns emerge. The first of these is that FLIPs appear to have greater overall sensitivity to Ala substitution than FunCs (Figures 2.3, 2.4). FLIPs have strong positive correlations with Sum  $\Delta\Delta G$ , #hot, and Avg  $\Delta\Delta G$  in PC1, while FunCs are negatively correlated with these traits (Figure 2.3b,c and Figure 2.4). This suggests the FLIPs have more specific interactions that produce large disruptions on Ala substitution than those of FunCs, a finding that agrees with experimental work [1] and is consistent with the characterization of weak and strong interfaces [7].

FLIPs also appear to have larger magnitude feature correlations along PC2 than FunCs, which cluster closer to the PC2 origin. PC2 is dominated by Slope\_ΔΔG, intcpt\_ΔΔG, and #total (Figure 2.3b,c), all 3 of which are statistically distinct between FLIP and FunC ( $P \leq 0.09$ ,  $P \leq 0.0006$ ,  $P \leq 0.0001$ ). Taken together, the correlations along PC2 suggest FLIPs have a strong central tendency with hotter centers and more interfacial residues than FunCs. This central tendency of FLIPs is also shown in Figure 2.1. While superficially, this is in agreement with certain precepts of Bogan and Thorn's "O-ring" hypothesis [8], it helps explain failures of the O-ring hypothesis to explain confounding examples of structures with hydrophilic or mixed hydrophilic and hydrophobic interfaces. A central tendency towards stability could be present in both proteins that follow a hydrophobic O-ring type structure, but could also be present in more hydrophilic interfaces that rely more on solvent and electrostatic interactions.

#### Implications for Interface Evolution

The emergence of both a larger specificity and a central organizing tendency from our ECR methodology suggests a model of interface evolution in which nascent, fortuitous interactions in a loose protein-protein association develop residue contacts that improve biological function for the organism. These interactions may have a selective pressure to be maintained or even improved (via mutation) in order to maintain or enhance the specific affinity of the two protein chains (Figure 2.2c-f). Residues surrounding these contacts may also have pressure to enhance affinity. Over evolutionary time, these selective pressures on the size and specific affinity produce a radially

symmetric pattern in the energetics of the interface (Figure 2.2b,f). The resulting interface should demonstrate “stronger” energies near the “older” regions of the interface. This hypothesis qualitatively agrees with the Evolutionary Trace results of Lichtarge and colleagues, who identify radially symmetric “bulls-eye” sequence conservation patterns near functionally important residues [49]. It also helps explain why sequence conservation methods alone without spatial, accessibility, or energetic contributions do not perform well as PPI predictors [24]. As the selective pressure on an interface is on energetic affinity and specificity, not sequence identity, FLIP interfacial residues may actually demonstrate larger sequence variation during the evolutionary “optimization” events. This can occur since improvements in specific affinity could arise if residues in both sides of a PPI were replaced via mutation. Similarly, one would not expect interfaces that are not acted upon by natural selection to have *a priori* central tendency patterns (Figure 2.2f). They should instead show a more random distribution of important residues (Figure 2.2c,d).

The ECR concept that evolutionary pressure will produce central tendency patterns with large specificity helps explain some discrepancies in our PCA/K-means cluster data as well. Both Antibody-Antigen and Inhibitor sub-categories cluster near the FunCs and XFunCs in our analysis (Figure 2.4). While antibody-antigen interactions are decidedly functionally linked, their quaternary structures are generally not evolutionarily driven. Instead, they are produced in a stochastic manner during V(D)J recombination [50]. As somatic cell hypermutation and B-cell selection is an

evolutionary-like process [51] and antibody-antigens are minimally oligo-trimers, it is also likely that center of interface of a large oligomer is not near the pairwise center, thus obscuring any central tendency. Similarly, enzyme inhibitors are often produced by infectious organisms to impede a host's native functions. While, the infecting organism may have a selective pressure to improve inhibitor binding, the host organism actually has selective pressures to escape inhibitor binding. For both antibodies and inhibitors, the lack of a *pairwise* central organizing tendency is thus not unlikely and may explain why these two functional sub-categories cluster with the FunCs.

### Implications for Protein Docking

Many protein-docking methods attempt to determine PPI structures by rapidly identifying and scoring regions of complementary shape and electrostatics [38]. Unfortunately, the large false positive rates of most docking methods reduce the usefulness of docking approaches [38, 41, 42]. Presumably, docking calculations are identifying regions of quaternary interaction conformational space that are not accessed by native conformations. As ECR can successfully distinguish FLIP conformations from FunC conformations, we propose ECR's use as a post-filter on the poses resulting from docking calculations. Our preliminary attempts at this look promising. As a proof of concept, we filtered the top 500 scoring poses generated by the docking program Hex [52] with ECR for several Enzymes and Antibody-HL interactions (1tzi\_AB, 1bsr\_AB, 1bsl\_AB, 1biq\_AB). In all these, we were able to identify the lowest RMSD

pose and in one case, 1bsr\_AB, were able to identify a lower RMSD pose than Hex.

Though very preliminary, we expect that our ECR method may substantially reduce false positive rates.

## **Conclusions**

In this work, we have introduced the FLIPdb, a database of protein-protein interfaces categorized by biological function. We have also introduced the Energy Centrality Relationship (ECR) method for analysis of computational alanine scan energetic distributions within protein-protein interfaces. We have successfully identified energetic and geometric features of interfaces that may be used to distinguish between functionally-linked (FLIP) and functionally uncorrelated (FunC) interfaces with a 76-80% accuracy. We have identified that FLIP interfaces have a stronger specificity and central organizing tendency than FunCs. Our ECR model also successfully describes the impact of evolution and natural selection on protein-protein interfaces. Finally, our ECR method may be of use in reducing the false positive rate of docking calculations.

## **Methods**

Dataset: FLIPdb

We collected a set of atomic-resolution structures all of which are available in the PDB [30] and then used additional literature and database sources to manually assign protein-protein interfaces to pre-decided categories. The database consists of 94 structures involving 233 individual proteins chains that formed 160 interfaces, which

were grouped into two primary categories, functionally-linked (FLIP) or functional uncorrelated (FunC). We initially combined selected subsets of structures from the databases of Janin and Weng [7, 43-45]. These datasets characterize proteins by whether they are known to be in protein complexes, have crystal contacts, are weakly or strongly interacting in solution, and how difficult they are to predict. Finally, we supplemented these with structures of general interest in our research. In this work, we chose to expand from prior datasets rather than simply use the datasets outright as these other sets were created to study specific questions but more importantly, did not always clearly delineate biological functional relevance of the PPI. For this work, we limited our selections to only bound complexes in an effort to purposefully limit structural variability and thus bias towards conformations with enhanced specificity. From this initial set, structures with resolutions greater than 3Å were rejected. We also generally excluded structures with very large cavities or projections whose curvature would produce interface centroids (based on C $\alpha$  positions) either out in space or far within the interior of one of the binding partners. We further removed any structure with 2 or fewer residues in the interface, partly in an effort to bias towards larger affinities and partly because the use of linear regression to map geometric features requires at least 3 bodies. We rejected structures with disordered residues or heteroatoms other than water or simple ions in the interface in order to bias the analyses towards amino acid interactions.

For each of the resultant interfaces, we performed a limited literature search focused on identifying: (1) whether the proteins were known to oligomerize *in vivo*; (2)

whether the proteins were known to oligomerize *in vitro* but under conditions similar to those within living systems; and (3) if mutations, post-translational modification, chemical modification, or small-molecule binding of residues within the interface were known to alter the function of the protein. (4) Additionally, we identified PPIs whose quaternary geometries were generally indicative of biological function, such as cytoskeletal proteins, viral capsid proteins, or immunoglobulin interactions between the heavy-chains as well as immunoglobulin heavy-chain:light-chain interactions outside the Fv region. We noted, but did not exclusively depend upon, whether PDB/PISA had designated the interface as being present in a Biological Assembly Unit (BioUnit). We categorized interfaces passing all these tests as FLIP. In addition, as a tool to aid our categorization, we noted whether the proteins could be simplistically sub-categorized into: (1) antibody-antigen (AbAg); (2) immunoglobulin Heavy Chain/Light Chain (AbHL); (3) Enzyme-Enzyme, both transient and persistent (Enzyme); (4) having a generally persistent structure that provides mechanical stability, such as cytoskeletal or viral proteins (Structural); (5) peptide/protein inhibitors to an enzyme (Inhibitor); (6) proteins whose function is to recognize peptides/proteins (Receptor); or (7) proteins regulated by post-translational modification by another protein (Regulated). We elected to use these 7, admittedly simplistic, operationally-defined sub-categories, rather than use SCOP [53], CATH [54], or GO [55] designations in order to limit the number of sub-categories and thus examine general FLIP characteristics. This is also consistent with categorizing all PPI into only the 2 FLIP/FunC categories. Most interfaces that could not be annotated as

FLIP were categorized as FunC, though some interfaces were eliminated from study if a number of conflicting annotations existed.

As the exclusions mentioned previously tended to eliminate FunC structures, we augmented our FunC numbers in two ways. First, we increased the number of proteins with a functionally unrelated PPI in the asymmetric unit by following the inverse of the method of Dey et al. [7]. We supplemented our set with entries from the PiQSi server [37] that were listed as solution-state monomers yet also had an entry of “PROBYES” in the Error field that indicates whether literature is in conflict with the reported quaternary assessment at PDB/PISA. Secondly, we utilized the available crystal symmetries to transform the coordinates of FLIP proteins such that crystal packing contact interfaces were produced. These transformations were created using the SYMEXP module of Pymol [56] and were sub-categorized as XFunCs. While it is generally desirable to have low similarity between dataset members to minimize compositional bias, our use of XFunCs derived from FLIPs actually provides a valuable internal control in that the two should be distinguishable. Failure to distinguish XFunCs from FLIPs in the same protein might suggest that general features of the protein rather than the interface were being biased towards. In order to further increase our FunC structures while maintaining some continuity with the datasets from the literature, we also created XFunCs from a subset of the members of the weakly interacting set of Dey et al. that were listed as only having crystal symmetry. All additional FunCs/XFunCs were also rejected if they failed to pass the same exclusionary limits placed on existing FLIPs

and FunCs. In addition, we rejected XFunC structures that literature review suggested might in reality be FLIP. The final database consisted of 94 structures comprised of 219 individual proteins chains that formed 160 interfaces. Of these, 100 were FLIP interfaces and 60 were FunC interfaces. Summary statistics of the FLIPdb are shown in Table 2.1.

In addition to this training set of interfaces, 18 additional interfaces (Test-18) were analyzed in order to provide a test set for cross-validation. All but two of the proteins in Test-18 had less than 70% sequence identity to proteins in the training set. Identity was determined using BLAST [57] run with default parameters available at servers at the National Center for Biotechnology Information. The remaining 2 proteins (immunoglobulin chains), though not identical to immunoglobulins in the training set, did have substantial similarity outside of the Fv region. These 18 PPI were subjected to the same physical and literature exclusionary limits as the training set. Summary statistics of the Test-18 are shown in Table 2.1.

Finally, as the training set had 38 members in common with the set of Dey and colleagues (16 weak and 22 strong in the training set), we created a second cross-validation testing set from 32 additional weak and 138 additional strong interfaces of Dey and colleagues [7]. Dey and colleagues purposefully characterized PPI predicted to have some level of oligomerization in solution, some weakly but most strongly. It is tempting to presume that the majority of these proteins would have some functional importance since they oligomerize in solution. However, without literature curation, one can only assume either (a) that all 170 PPI are FLIP or (b) the strong PPI are more FLIP-like and

the weak PPI are more FUNC-like. These assumptions were evaluated in this work.

Summary descriptions of these 170 PPI are listed in Table S2.2.

### Computational Alanine Scanning (CAS)

The CAS method of Kortemme and Baker [27, 28, 58], was used to process all the interfaces in the FLIPdb. In brief, this method evaluates enthalpy and free energy of solvation terms over conformations arising from a rotamer library for both the existing and alanine substituted residues in a PPI (native Gly and Pro excluded). These terms are used to determine a pseudo-free energy change upon substitution ( $\Delta\Delta G$ ) [28].

Computational Alanine Scanning (CAS) calculations were performed using the Agnito HPC Linux cluster at Texas Woman's University according to scripts and libraries kindly supplied by Dr. Tanja Kortemme (UCSF). These results were spot-checked against CAS calculations made using the ROBETTA server of David Baker's lab [59]. In all cases the results were identical.

### Interfacial Geometry

Interfacial residues were defined using the same interface definition as in the CAS method of Kortemme and Baker [28]. The geometric distribution of residues in each PPI were determined by calculating the displacement ( $\Delta r$ ) of the  $C\alpha$  position from the mean of the  $C\alpha$  positions (termed the Center of Interface, CoI) using software written by the authors. A linear regression of the  $\Delta\Delta G$  and  $\Delta r$  data to a first-order polynomial ( $\Delta\Delta G = \text{slope} * \Delta r + \text{intercept}$ ) was calculated for each interface using software written by

the authors as well as GNU PLOT [60]. The calculations provided 8 features for each interface: the slope (slope\_ΔΔG), intercept (intcpt\_ΔΔG), coefficient of determination (R<sup>2</sup>\_ΔΔG), net sum of all ΔΔG changes (sum\_ΔΔG), mean ΔΔG for all interface residues (avg\_ΔΔG), total number of residues in the interface (#total), number of residues with ΔΔG larger than +1 kcal/mol (#hot), and the ratio of “hot” to total (frac\_hot). Examples of the distribution of these ΔΔG values for a FLIP (PDBid: 1vfr) and FunC (PDBid: 1c02) are shown in Figure 2.1.

### Principle Component Analysis (PCA)

Principle component analysis of the variation of CAS energetic and geometric feature data for all PPI was undertaken using JMP [61]. PCA determines a set of linearly-uncoupled eigenvectors from normalized correlations between variables that progressively describe the largest sources of variance in a data set [46]. The eigenvector coefficients for each principal component vector indicate the relative correlation between each feature and the overall variation of all features. In this work, we sought to identify the set of features that would describe more than 80% of the total set variation in the first two principal components (PCs) such that we could use a minimum number of PCs to discriminate between FunC and FLIP data. The results from these PCA analyses are shown in Figures 2.3 and 2.4 and Table 2.2. Due to the lower contribution of the coefficient of determination (R<sup>2</sup>) of the linear regression towards overall feature variation, this term was dropped and only the remaining seven features were used.

## K-Means Clustering

K-means clustering [47] is a data analysis method that clusters observations into a specific number of clusters by attempting to find the point(s) that have the lowest mean variation from the other input data. When combined with PCA, the combination of features that allows input data to be clustered can be identified. In this work, two clusters were specified and the correlations between cluster and functional category determined (Figure 2.4a,b and Table 2.2). Forty-seven (47) FLIP interfaces (mostly enzyme and immunoglobulin heavy-chain/light chain interfaces) could easily be identified. A second round of PCA and K-means clustering excluding these 47 FLIP (and 2 FunC PPI falsely identified as FLIP) was subsequently performed (Figure 2.4c,d and Table 2.2).

## Accuracy and Matthews Correlation Coefficient

The following measures were used to assess the performance of our clustering analysis:

*Accuracy* (ACC), the propensity to correctly identify FLIP or FunC:

$$ACC = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

and *Matthews correlation coefficient* (MCC), a measure of how much a set of predictive data agrees with a two-state model:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (2)$$

where,

TP = the number of interfaces correctly predicted as FLIPs (True Positive)

TN= the number of interfaces correctly predicted as FunCs (True Negative)

FP = the number of interfaces wrongly predicted as FLIPs (False positive)

FN = the number of interfaces wrongly predicted as FunCs (False Negative)

These values are shown in Tables 2.2, S2.3.

### **Acknowledgments**

We thank Texas Woman's University Office of Technology for high performance computing resources. We would like to acknowledge the valuable discussions Drs. Lynda Uphouse and DiAnna Hynds (both TWU).

## References

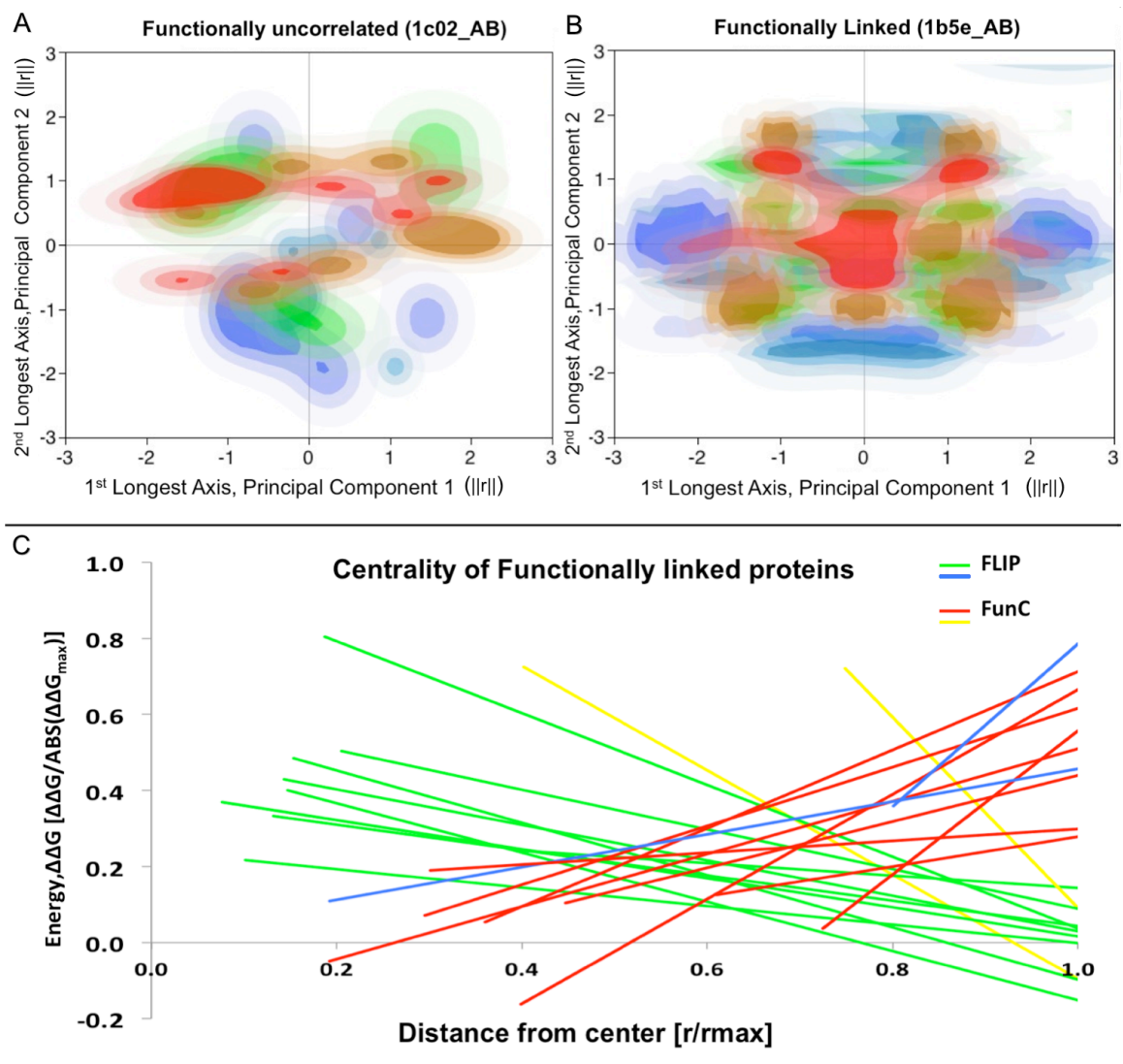
1. Phizicky EM, Fields S. (1995) Protein-protein interactions: Methods for detection and analysis. *Microbiol Rev* 59: 94-123.
2. Wells CT. (1995) A hot spot of binding energy in a hormone-receptor interface. *Science* 267: 383-386.
3. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. (2000) A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae*. *Nature* 403: 623-627.
4. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. (2002) Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectrometry. *Nature* 415: 180-183.
5. Gavin A, Bösche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415: 141-147.
6. Bahadur RP, Zacharias M. (2008) The interface of protein-protein complexes: Analysis of contacts and prediction of interactions. *Cell Mol Life Sci* 65: 1059-1072.
7. Dey S, Pal A, Chakrabarti P, Janin J. (2010) The subunit interfaces of weakly associated homodimeric proteins. *J Mol Biol* 398: 146-160.  
10.1016/j.jmb.2010.02.020.
8. Bogan AA, Thorn KS. (1998) Anatomy of hot spots in protein interfaces. *J Mol Biol* 280: 1-9.
9. Valencia A, Pazos F. (2002) Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol* 12: 368-373.
10. Ofra Y, Rost B. (2003) Analysing six types of protein-protein interfaces. *J Mol Biol* 325: 377-387.
11. Zhu H, Domingues FS, Sommer I, Lengauer T. (2006) NOXclass: Prediction of protein-protein interaction types. *BMC Bioinformatics* 7: 27-27.
12. Duarte JM, Srebniak A, Schärer M, A., Capitani G. (2012) Protein interface classification by evolutionary analysis. *BMC Bioinformatics* 13: 334-334.  
10.1186/1471-2105-13-334.

13. Krissinel E, Henrick K. (2005) Detection of protein assemblies in crystals. In: R. Berthold M, Glen R, Diederichs K, Kohlbacher O, Fischer I, editors. : Springer Berlin Heidelberg. pp. 163-174. 10.1007/11560500\_15.
14. Young KH. (1998) Yeast two-hybrid: So many interactions, (in) so little time.. *Biol Reprod* 58: 302-311.
15. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, et al. (2001) BIND--the biomolecular interaction network database. *Nucleic Acids Res* 29: 242-245.
16. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, et al. (2002) MINT: A molecular INTeraction database. *FEBS Lett* 513: 135-140.
17. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, et al. (2002) DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* 30: 303-305.
18. Valdar WS, Thornton JM. (2001) Protein-protein interfaces: Analysis of amino acid conservation in homodimers. *Proteins* 42: 108-124.
19. Fleishman SJ, Whitehead TA, Strauch E, Corn JE, Qin S, et al. (2011) Community-wide assessment of protein-interface modeling suggests improvements to design methodology. *J Mol Biol* 414: 289-302. 10.1016/j.jmb.2011.09.031.
20. Bahadur RP, Chakrabarti P, Rodier F, Janin J. (2003) Dissecting subunit interfaces in homodimeric proteins. *Proteins* 53: 708-719.
21. Jones S, Thornton JM. (1996) Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* 93: 13-20.
22. DeLano WL. (2002) Unraveling hot spots in binding interfaces: Progress and challenges. *Curr Opin Struct Biol* 12: 14-20.
23. Liu S, Li Q, Lai L. (2006) A combinatorial score to distinguish biological and nonbiological protein-protein interfaces. *Proteins* 64: 68-78.
24. Ofra Y, Rost B. (2007) Protein-protein interaction hotspots carved into sequences. *PLoS Comput Biol* 3: e119-e119.
25. Massova I, Kollman PA. (1999) Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *J Am Chem Soc* 121: 8133-8143. 10.1021/ja990935j.

26. Thorn KS, Bogan AA. (2001) ASEdb: A database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* 17: 284-285. 10.1093/bioinformatics/17.3.284.
27. Kortemme T, Baker D. (2002) A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A* 99: 14116-14121.
28. Kim KT. (2004) Computational alanine scanning of protein-protein interfaces. *Sci STKE* 2004: pl2-pl2.
29. Meenan NAG, Sharma A, Fleishman SJ, Macdonald CJ, Morel B, et al. (2010) The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proc Natl Acad Sci U S A* 107: 10080-10085. 10.1073/pnas.0910756107.
30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. *Nucleic Acids Res* 28: 235-242.
31. Henrick KE. (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 372: 774-797.
32. Krissinel E. (2010) Crystal contacts as nature's docking solutions. *J Comput Chem* 31: 133-143.
33. De Las Rivas J, Fontanillo C. (2010) Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Comput Biol* 6: e1000807. 10.1371/journal.pcbi.1000807.
34. Nooren IM, Thornton JM. (2003) Diversity of protein-protein interactions. *EMBO J* 22: 3486-3492.
35. Henrick PH. (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins* 41: 47-57.
36. Havranek JJ. (2010) Specificity in computational protein design. *J Biol Chem* 285: 31095-31099. 10.1074/jbc.R110.157685.
37. Levy ED. (2007) PiQSi: Protein quaternary structure investigation. *Structure* 15: 1364-1367.
38. Janin J. (2005) Assessing predictions of protein-protein interaction: The CAPRI experiment. *Protein Sci* 14: 278-283.

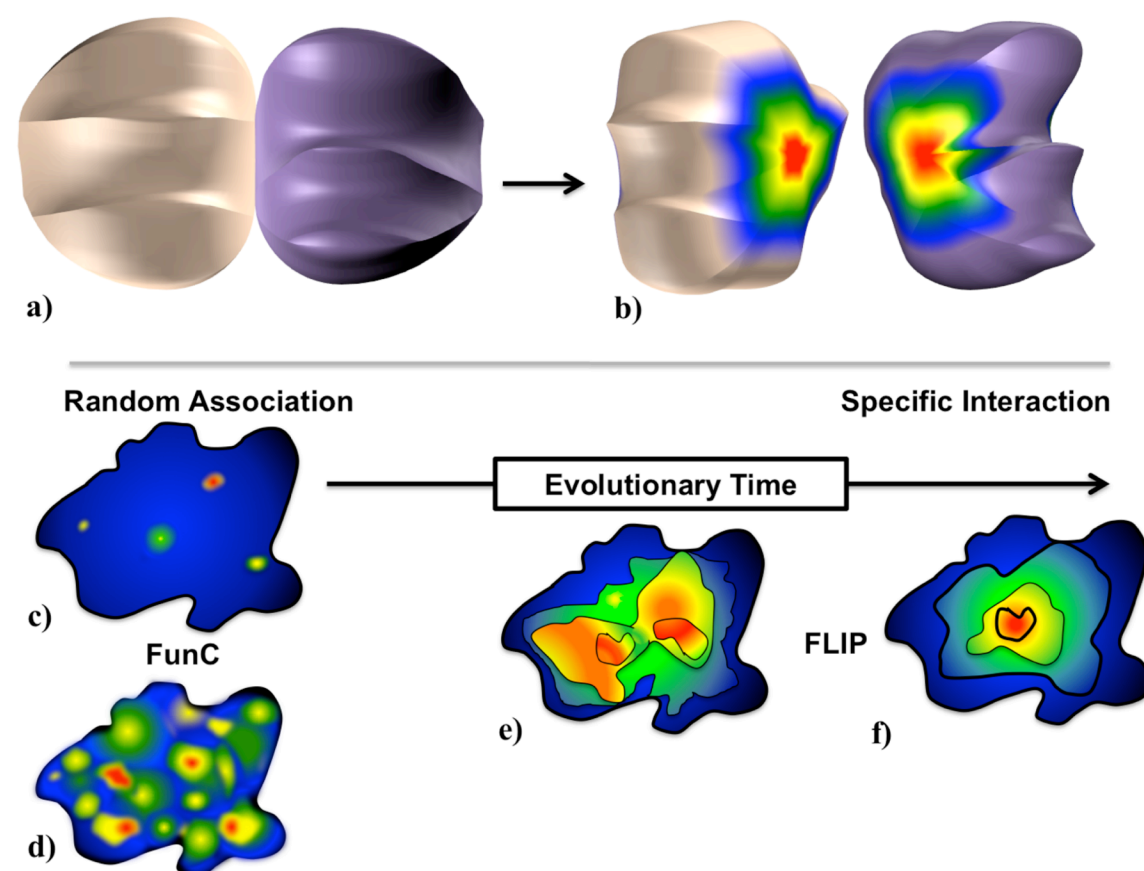
39. Janin J. (2010) Protein-protein docking tested in blind predictions: The CAPRI experiment. *Mol Biosyst* 6: 2351-2362. 10.1039/c005060c.
40. Moretti R, Fleishman SJ, Agius R, Torchala M, Bates PA, et al. (2013) Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. *Proteins* .
41. Shoichet BK, McGovern SL, Wei B, Irwin JJ. (2002) Lead discovery using molecular docking. *Curr Opin Chem Biol* 6: 439-446.
42. Aytuna AS, Gursoy A, Keskin O. (2005) Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics* 21: 2850-2855.
43. Bahadur RP, Chakrabarti P, Rodier F, Janin J. (2004) A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol* 336: 943-955.
44. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, et al. (2005) Protein-protein docking benchmark 2.0: An update. *Proteins* 60: 214-216.
45. Chakrabarti P, Janin J. (2002) Dissecting protein-protein recognition sites. *Proteins* 47: 334-343.
46. Pearson K. (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2 11: 559-572.
47. Hartigan JA. (1973) Clustering. *Annu Rev Biophys Bioeng* 2: 81-101. 10.1146/annurev.bb.02.060173.000501.
48. Dey S, Pal A, Chakrabarti P, Janin J. (2010) The subunit interfaces of weakly associated homodimeric proteins. *J Mol Biol* 398: 146-160. 10.1016/j.jmb.2010.02.020.
49. Mihalek I, Reš I, Lichtarge O. (2004) A family of Evolution–Entropy hybrid methods for ranking protein residues by importance. *J Mol Biol* 336: 1265. 10.1016/j.jmb.2003.12.078.
50. Schatz DG, Swanson PC. (2011) V(D)J recombination: Mechanisms of initiation. *Annu Rev Genet* 45: 167-202. 10.1146/annurev-genet-110410-132552.
51. Schwartz GW, Hershberg U. (2013) Germline amino acid diversity in B cell receptors is a good predictor of somatic selection pressures. *Front Immunol* 4: 357-357. 10.3389/fimmu.2013.00357.

52. Ritchie DW, Kemp GJ. (2000) Protein docking using spherical polar fourier correlations. *Proteins* 39: 178-194.
53. Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C. (1999) SCOP: A structural classification of proteins database. *Nucleic Acids Res* 27: 254-256.
54. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, et al. (1997) CATH--a hierarchic classification of protein domain structures. *Structure* 5: 1093-1108.
55. Ball AM. (2000) Gene ontology: Tool for the unification of biology. the gene ontology consortium. *Nat Genet* 25: 25-29.
56. Schrödinger L. The PyMOL molecular graphics system. 1.5.0.4.
57. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. *J Mol Biol* 215: 403-410.
58. [Anonymous]. [Http://Robetta.bakerlab.org](http://Robetta.bakerlab.org). .
59. Kim DE, Chivian D, Baker D. (2004) Protein structure prediction and analysis using the robetta server. *Nucleic Acids Res* 32: W526-W531.
60. Williams T, Kelley C. (2011) .
61. SAS Institute Inc., Cary, NC. (1989-2012) JMP. 10.



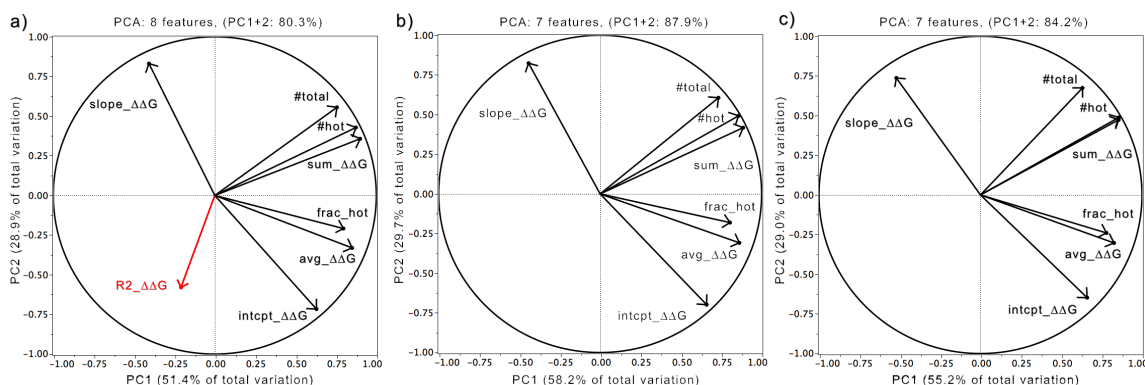
**Figure 2.1. Distribution of Alanine Substitution Energies in FLIP and FunC Interfaces.**

**Figure 2.1. Distribution of Alanine Substitution Energies in FLIP and FunC Interfaces.** (a) and (b) show a histogrammed contour plot colored blue-to-red of the  $\Delta\Delta G_{\text{ala}}$  of substitution to alanine of interfacial residues (blue: more favorable values, red: more disruptive values). The plot axes are the first two principal components of the geometric distribution of alanine C $\alpha$  positions. PCA was used to align the interface along the X- and Y-axes. Axes are normalized. (a)  $\Delta\Delta G_{\text{ala}}$  of the FunC interface from PDBid: 1c02, chains A&B. (b)  $\Delta\Delta G_{\text{ala}}$  of the FLIP interface from PDBid: 1b5e\_AB, chains A&B. (c) Linear regressions of  $\Delta\Delta G_{\text{ala}}$  vs. Distance from interface center. Regressions for the interfaces in the FLIPdb training set with the 10 most positive [1acy\_HP, 1biq\_AB, 2cii\_AC, 1b5e\_AB, 1edh\_AB, 1pky\_BD, 1tx4\_AB, 1hjc\_AC, x1bsf8\_AJ, 1bo5\_OZ] and 10 most negative [1tzi\_AV, 1acy\_LP, x1ppf2\_EZ, x1dv82\_AC, x1wtl\_BZ, x1py94\_AE, x1erv2\_AC, x1gaf2\_LY, 1scu\_BD, 1c02\_AB] intercepts. FLIP are shown in green and blue [1tzi\_AV, 1acy\_LP]. FunC are shown in red and yellow [x1bsf8\_AJ, 1bo5\_OZ].  $\Delta\Delta G_{\text{ala}}$  are normalized to  $\text{MAX}(\text{ABS}(\Delta\Delta G_{\text{ala}}))$ , while distances of each residue's C $\alpha$  from the mean of the C $\alpha$  positions (Center of Interface) are normalized to  $\text{MAX}(\text{distance})$ . All 3 plots generally show that FLIP interfaces are more centralized and radially symmetric than FunC interfaces. 80% of shown positive intercepts are FLIP and 80% of shown negative intercepts are FunC. [Figures (a,b) generated using JMP [46]. Figure (c) generated using Microsoft Excel, 2008]

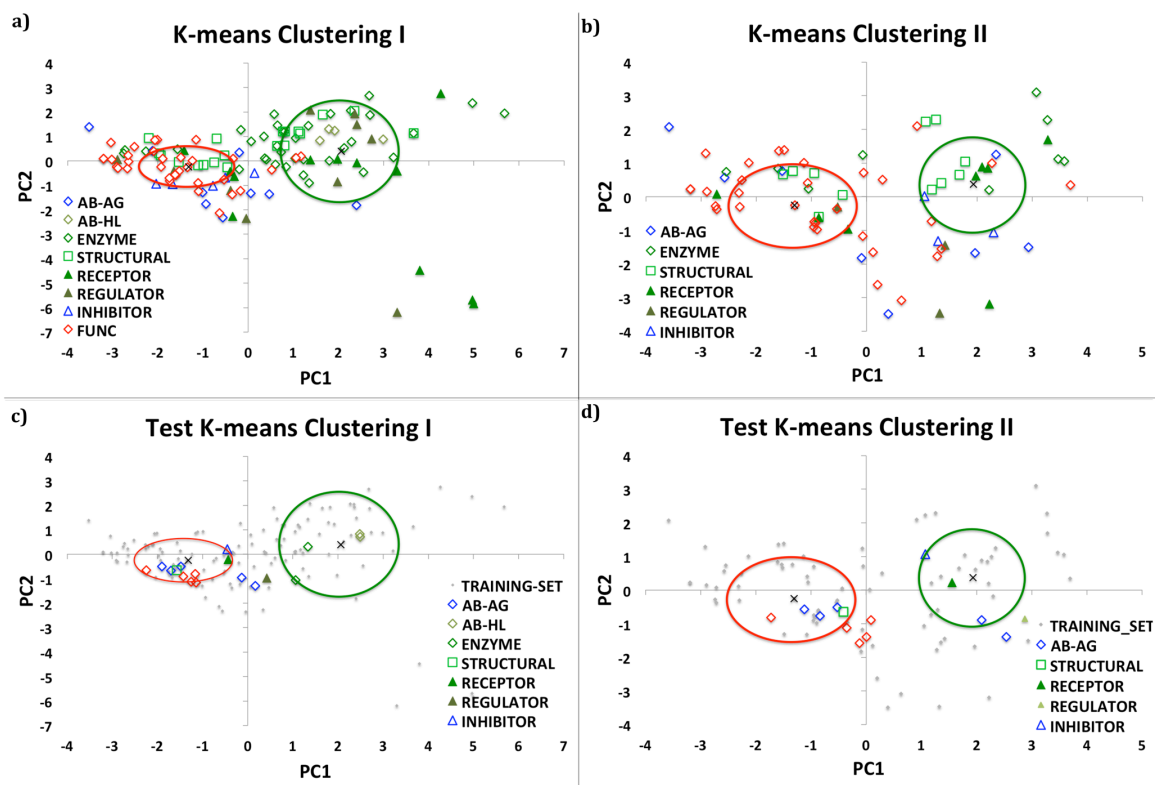


**Figure 2.2. The Energy Centrality Relationship (ECR) for Interface Evolution.**

The ECR hypothesis is that upon initial fortuitous protein-protein association, residues in a nascent interface have a selective pressure to maintain or improve the affinity arising from the initial contact, while simultaneously having a similar pressure on residues surrounding that contact. (a) and (b) show a conceptual PPI that has a radially symmetric distribution of ‘hot’ (energetically favorable, red) and ‘cold’ (energetically unfavorable, blue) residues in a FLIP, while (c) and (d) are example residue energy distributions of weaker (c) and stronger (d) affinity FunC. Over evolutionary time (c-f), selective activity, affinity, and specificity pressures on residues in a FunC produce a radially symmetric pattern in the energetics of the interface. The resulting interface should demonstrate “stronger” energies near the “older” regions of the interface. These “older” regions may or may not demonstrate sequence conservation as the pressure is on energy, not identity. As natural interfaces are generally more punctate than the ideal model, we expect that while both FLIP and FunC interfaces may demonstrate multiple contacts, only FLIP interfaces will maintain overall centrality (e-f).



**Figure 2.3. Correlation of Features with Principal Components.** Loading plots of the eigenvector coefficients of each feature analyzed by PCA show the influence and correlations of each variable to the principal components. Eight features were analyzed to identify the set of features that could represent ~80% of data variation in the first two principal components (see text for feature descriptions). (a) 80.3% of the total variance of all eight features could be accounted for with just the first two PCs, though R2\_ΔΔG (red) had demonstrably smaller coefficients. (b) Exclusion of R2\_ΔΔG produced a PCA over 7 features whose PC1 and PC2 accounted for 87.9% of the variance. (c) After removal of 49 interfaces predicted to be FLIP in the first PCA, a second round of PCA using the same seven features but with only data for the remaining 110 protein interfaces was calculated. This PCA produced eigenvectors that had 84.2% of the variance in the first two PCs. [Figure generated using JMP [46] and Microsoft Excel, 2008].



**Figure 2.4. PCA and K-means Clustering of Training and Test-18 Sets.** Principal component analysis followed by K-means clustering was performed on the residues in the 100 FLIP and 60 FunC interfaces in the FLIPdb. The same 7 features identified in Figure 2.3 are used here and the number of clusters was set to  $k=2$ . Green (“cluster 1”) and red (“cluster 2”) ovals represent 1 standard deviation for Euclidean distances around the cluster centroid marked by “x”. Interfaces are indicated with symbols representing their functional sub-category. Green and Blue symbols are FLIP structures, but blue symbols are specifically AbAg and Inhibitor sub-categories. Red symbols are FunCs. (a) and (b): training set. (c) and (d): Test-18 testing set. (a) 49 FLIP interfaces (mostly enzymes and immunoglobulin Heavy-Light chains) and 1 FunC are identified in cluster 1 (98% precision). (b) After removal of these 50 interfaces, a second PCA analysis of the remaining 110 interfaces produces new clusters with 48 and 62 members, respectively. PCA 2 Cluster 1 is 64% FLIP and cluster 2 is 68% FunC. Overall accuracy across both (a)+(b) is 76%. (c) and (d) show the projection of the 7 feature values 18 unrelated PPIs in the Test-18 set through the principal components developed on the training set. Enzymes and immunoglobulin Heavy-Light again dominate cluster 1 (100%) and overall accuracy in both clusterings is 78%. [Figure generated with JMP [46] and Microsoft Excel, 2008]

**Table 2.1.** Summary of Protein and Protein Interface Counts in FLIPdb.

Function		Training Set			Test 18 Set		
Category	Sub-categories	PDB Structures	Protein chains	Interfaces	PDB Structures	Protein chains	Interfaces
FLIP	AbAg*	4	15	12	1	6	5
	AbHL*	5	10	5	1	4	2
	Enzyme	33	74	40	2	4	2
	Structural	7	21	16	1	2	1
	Receptor	7	16	10	1	2	1
	Regulator	9	20	12	1	2	1
	Inhibitor	3	10	5	1	2	1
	Total	63	155	100	7	18	13
FunC	FunC	22	47	25	-	-	-
	XFunC‡	23	44	35	5	10	5
	Total	44	89	60	5	10	5
Total		94	219	160	7	19	18

\* Proteins chains are common to multiple sub-categories though the interfaces are distinct.

‡ Interfaces are constructed from existing FLIPs through coordinate transformations arising from the symmetry of the source X-ray crystal structure (XFunCs).

FLIPdb contains 160 interfaces in 94 structures involving 219 individual protein chains. These interfaces have been assigned to FLIP or FunC functional categories and 9 functional sub-categories based on a review of the literature (see Supplement Table S2.1). Due to the reuse of some chains, the totals represented in the first two columns do not sum across sub-categories.

**Table 2.2.** Accuracy of Clustering in Training and Test-18 Sets.

	True Positive (TP)	False Positive (FP)	False Negative (FN)	True Negative (TN)	Accuracy	MCC
Training Set†	1 <sup>st</sup> clustering					
	49	1	51	59	67.5%	0.49
	2 <sup>nd</sup> Clustering					
	31	17	20	42	66.4%	0.32
Total*	80	18	20	42	76.3%	0.50
Test 18 Set†	1 <sup>st</sup> clustering					
	3	0	10	5	44.4%	0.28
	2 <sup>nd</sup> clustering					
	6	0	4	5	73.3%	0.58
Total*	9	0	4	5	77.8%	0.62

†TP: FLIP found in Cluster 1

†TN: FUNC found in Cluster 2

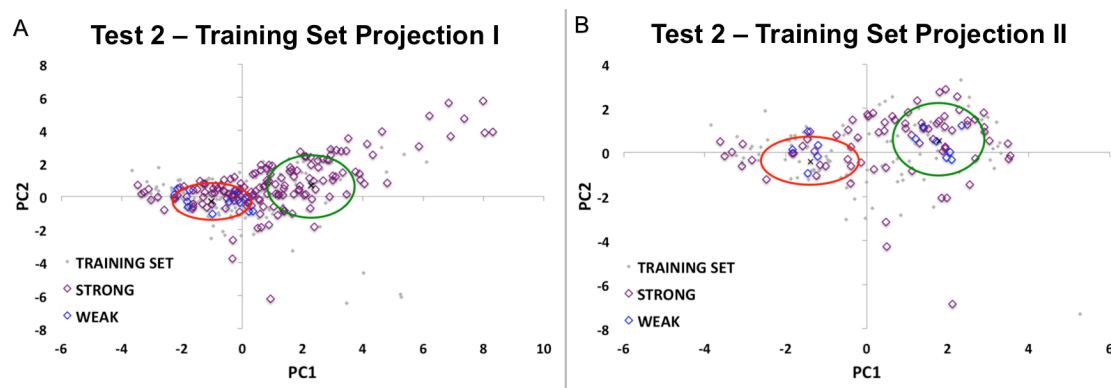
†FP: FUNC found in Cluster 1

†FN: FLIP found in Cluster 2

\*Total FNs and TNs do not add sum across 1<sup>st</sup> and 2<sup>nd</sup> clustering. TP and FP sum, but FN and TN arise only from 2<sup>nd</sup> clustering

The accuracy and Matthews correlation coefficient (MCC, a measure of the quality of a binary classification) of the results of the clusterings shown in Figure 2.4 are indicated. The overall accuracy is 76% and 78% for both training Test-18 sets, respectively. TPs are quite readily identified in both training and Test-18 sets (80% and 69% *sensitivity*, respectively). The majority of TPs are enzymes and immunoglobulin heavy chain-light chain interactions. TNs are less well identified (70% and 56% *negative predictive values*, respectively). MCCs of 0.50 and 0.62 indicate that our simple two-category approach is generally appropriate.

## Supporting Information



**Figure S2.1. PCA and K-means Clustering of the Dey-170 Set.**

Projection of the 7 feature values of the PPI in the Dey-170 set through the principal components developed on the training set. Grey dots show the values of the training set. Green and red ovals represent 1 standard deviation for Euclidean distances around the cluster centroid marked by “x”. Values for Dey-170 interfaces are indicated with purple symbols representing “Strong” PPI interactions and blue symbols representing “Weak” PPI interactions. (a) and (b) shows projections through PCA 1 and 2 principal components, respectively. (a) 60% of Strong PPI and 0% Weak PPI group in cluster 1 while 40% of Strong and 100% of Weak group in cluster 2, yielding 100% *precision* and 100% *negative predictive value*. (b) After removal of the 82 PPI in cluster 1, a second projection of the 88 remaining values through PCA 2 produces new clusters with 54 and 34 members, respectively. PCA 2 Cluster 1 is 78% Strong while cluster 2 is 59% Weak. [Figure generated with JMP [46] and Microsoft Excel, 2008]

## Supplementary Tables

Due to dissertation formatting constraints, please see the following webpage for full table: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0097115>

### **Table S2.1.** FLIPdb Interface Composition

Structures and interfaces used in the training and testing sets. The FLIPdb database contained 160 pairwise PPI between 219 protein chains that were contained in 94 PDB structural files. The Test-18 set contains 18 pairwise PPI between 19 proteins chains contained in 7 PDB files. Based on literature review, these PPIs were categorized into the FLIP or FunC interface class (100 FLIP, 60 FunC). The PPIs were further sub-categorized into 7 FLIP and 2 FunC sub-categories: 1) antibody-antigen (AbAg); 2) immunoglobulin Heavy Chain/Light Chain (AbHL); 3) Enzyme-Enzyme, both transient and persistent (Enzyme); 4) having a generally persistent structure that provides mechanical stability, such as cytoskeletal or viral proteins (Structural); 5) peptide/protein inhibitors to an enzyme (Inhibitor); 6) proteins whose function is to recognize peptides/proteins (Receptor); 7) proteins regulated by post-translational modification by another protein (Regulated); 8) PPIs in an asymmetric crystal unit NOT found to be FLIP (FunC); and 9) PPIs obtained by applying crystal symmetry transforms to FLIP structures (XFunC). The Dey-170 set contains 170 pairwise PPI between 301 proteins chains contained in 139 PDB files. Categories were uncured and sub-categories of “Strong” and “Weak” were derived from [7]. The number of chains, number of interfaces, and the references used to justify classification for each pairwise interface are listed.

**Table S2.2.** Summary of Protein and Protein Interface Counts in Dey-170.

Function		Dey-170 Set		
Category	Sub-categories	PDB Structures	Protein chains*	Interfaces
Unknown	Weak	17	46	32
	Strong	127	266	138
	Total	139	301	170

\* Proteins chains are common to multiple sub-categories though the interfaces are distinct.

Dey-170 contains 170 interfaces in 139 structures involving 301 individual protein chains. These interfaces have been labeled as Strong or Weak categories based on information available in [7] (see Supplement Table S2.1). Due to the reuse of some chains, the totals represented in the first two columns do not sum across sub-categories.

**Table S2.3.** Pseudo-Accuracy of Clustering in Dey-170 Test Set

<b>Dey-170 Testing Set</b>	<b>True Positive (TP)</b>	<b>False Positive (FP)</b>	<b>False Negative (FN)</b>	<b>True Negative (TN)</b>	<b>Accuracy</b>	<b>MCC</b>
a)† <sup>a</sup> Assuming All:FLIP	<b>1<sup>st</sup> clustering projection</b>					
	82	0	88	0	48.2%	-0.005
	<b>2<sup>nd</sup> clustering projection</b>					
	54	0	34	0	61.4%	0.05
<b>Total</b>	136	0	34	0	80.0%	0.12
b)‡ Assuming Weak:FUNC Strong:FLIP	<b>1<sup>st</sup> clustering projection</b>					
	82	0	56	32	67.1%	0.46
	<b>2<sup>nd</sup> clustering projection</b>					
	42	12	14	20	70.4%	0.37
<b>Total*</b>	124	12	14	20	84.7%	0.51

†TP: FLIP found in Cluster 1

†TN: FUNC found in Cluster 2

†FP: FUNC found in Cluster 1

†FN: FLIP found in Cluster 2

<sup>a</sup>MCC adjusted to set all zero-values to 1

‡TP: Strong found in Cluster 1

‡TN: Weak found in Cluster 2

‡FP: Weak found in Cluster 1

‡FN: Strong found in Cluster 2

\*Total FNs and TNs do not add sum across 1<sup>st</sup> and 2<sup>nd</sup> clustering. TP and FP sum, but FN and TN arise only from 2<sup>nd</sup> clustering

The accuracy and Matthews correlation coefficient of the results of the cluster projections for the PPI in the Dey-170 test set shown in Figure S2.1. Values likely represent a lower limit, as all PPI in this set were purposely not rigorously curated and were operationally presumed (a) to be FLIP or (b) for Weak to be FUNC and Strong interactions to be FLIP. (a) The overall accuracy is 80%. Subsequent projection rounds and the overall MCC shift positively to 0.12. The accuracy greater than 75% and MCC > 0 (b) Presuming Weak interactions match FUNC and Strong interactions match FLIP yields a larger accuracy of 84.7% and a MCC of 0.51.

**Table S2.4.** Random Sub-Sample Validation of FLIPdb Training Set

<b>% FLIPdb (# of Interfaces)</b>	<b>Trial</b>	<b>TP<sup>‡</sup></b>	<b>FP<sup>‡</sup></b>	<b>FN<sup>‡</sup></b>	<b>TN<sup>‡</sup></b>	<b>Accuracy</b>	<b>MCC</b>
<b>100</b>	1	80	18	20	42	76%	0.50
<b>90 (144)</b>	1	68	15	21	40	75%	0.48
	2	70	16	19	39	76%	0.49
	3	74	18	17	36	76%	0.48
<b>80 (128)</b>	1	67	19	11	31	77%	0.50
	2	64	15	16	33	76%	0.49
	3	64	14	19	30	74%	0.44
<b>70 (112)</b>	1	32	0	35	45	69%	0.52
	2	56	16	12	28	75%	0.47
	3	40	3	29	40	71%	0.51
<b>60 (96)</b>	1	39	4	22	31	73%	0.51
	2	47	15	12	22	72%	0.40
	3	53	14	12	17	73%	0.37
<b>50 (80)</b>	1	38	2	14	26	80%	0.63
	2	34	4	17	23	73%	0.49
	3	42	10	12	16	73%	0.39
<b>40 (64)</b>	1	15	3	20	26	64%	0.36
	2	30	12	8	14	69%	0.34
	3	25	4	14	21	72%	0.47
<b>30 (48)</b>	1	20	1	15	12	67%	0.44
	2	18	1	11	18	75%	0.57
	3	19	4	7	16	76%	0.53
<b>20 (32)</b>	1	15	2	4	11	81%	0.63
	2	12	5	7	8	63%	0.24
	3	10	0	15	7	53%	0.36

<sup>‡</sup> ) TP: Strong found in Cluster 1

TN: Weak found in Cluster 2

FP: Weak found in Cluster 1

FN: Strong found in Cluster 2

The distribution of overall accuracies and MCCs of repeated retraining when sub-samples of the training set were generated randomly in triplicate for subsets ranging from 90% to 20% of the original. The general accuracy is 70-80% until more than 50% of the training set is removed. MCCs stably range between approximately 0.20-0.60. This stability suggests little compositional bias in the FLIPdb training set.

CHAPTER III

FUNCTIONAL CLASSIFICATION OF PROTEIN INTERACTIONS USING  
INTERFACE SPATIAL DISTRIBUTION OF  
EVOLUTIONARY CRITERIA

A paper in revision at Protein Science

Sanjana Sudarshan<sup>1</sup>, Brian W. Beck<sup>1,2,3†</sup>

**Abstract**

Interacting networks of proteins carry out various biological functions such as providing structure, or carrying out metabolic or other cellular processes. Most contacts are identified via X-ray crystallography. Often, such interactions are simply crystal contacts that lack relevancy to the functions of those proteins. In previous work, we have used our Energy Centrality Relationship (ECR) concept to discriminate Functionally-Linked Interacting Protein (FLIP) structures from Functionally uncorrelated Contacts (FunC) by assessing how the energetic contribution of each interface amino acid varies as a function of distance from the interface. We hypothesize that evolutionary pressure plays an important role in maintaining FLIPs. Here we use Lichtarge's Evolutionary Trace (ET) method to calculate the ET score ( $\rho$ ) and alignment variability (# of states) of residues within various types of interfaces. Using Principal Component Analysis and K-means clustering on a set of 154 interfaces, we were able to distinguish FLIPs from

FunCs with an accuracy of 69%. Here we show that while having a central tendency and showing more conservation at the core than at the edges, FLIPs nevertheless had more variability than would otherwise be expected, perhaps demonstrating conservation of energy at the expense of sequence stability.

## **Introduction**

Protein-protein interactions (PPIs) form important components of cellular machinery and are involved in performing various biological processes<sup>1-5</sup>. Some key characteristic features of PPIs used in interface detection are shape and electrostatic complementarity, size of the interface (including buried surface area, number of residues, and number of core interface atoms), flexibility, solvation energy, and sequence conservation of the interactors (amino acid residues)<sup>6-12</sup>. Interfaces also often differ from the remainder of the protein surface, exhibiting the following characteristics: surface planarity<sup>13</sup>, enriched residue conservation<sup>14</sup>, prevalence of aromatic residues<sup>15</sup>, modular architecture<sup>16</sup>, uneven distribution of hot spots (energetically important residues)<sup>2, 17</sup>, and increased packing<sup>18</sup>. These features can be used to characterize PPIs and combinations of these features have been used to differentiate between: (a) biologically relevant interfaces and artifacts of crystallization<sup>19</sup>, (b) transient and obligatory interfaces<sup>9, 20</sup>, and (c) native and non-native poses in protein-protein docking<sup>21, 22</sup>. The presence or absence of these general characteristics have also led to various hypotheses in interface delineation; for example, one group suggested that non-conserved residues surround hot spots at the interface, commonly known as the O-ring hypothesis<sup>17</sup>. Another suggested the interface

could be divided into a core region, consisting of buried residues and a rim region, consisting of exposed residues<sup>23</sup>.

Even with these diverse sets of features available, in order to understand protein interactions, it is important to know which residues at the interface contribute the most to the stability of the interface<sup>24</sup>. X-ray crystallographic and NMR (Nuclear Magnetic Resonance) structures provide atomic level details of proteins but are often limited to individual domains that may also lack specific binding site information<sup>25</sup>. The importance of each residue towards maintenance of structure and function has often been assessed by looking at how conserved it is within a family of proteins<sup>8, 10, 26-30</sup>.

Residues critical to binding are more likely to be conserved<sup>12</sup> and located in structurally important locations and at active sites<sup>25, 28-32</sup>. In general, residues at interfaces are slightly more conserved when compared to surface residues<sup>26, 33</sup> but not significantly different from the interior<sup>34</sup>. This suggests that interfacial and core regions are important to maintaining structure and function, and that the surface residues are less actively involved in the folding and stability of the protein<sup>35</sup>. Conservation signals are observed by comparing interface to surface residues and a high signal-to-noise ratio could be indicative of mutations that are not well-tolerated<sup>35</sup>. It could be understood from evidence available that by coordinating with residues directly involved in binding, residues away from the binding sites could influence the stability of PPIs<sup>24</sup>.

It is possible conservation is seen at the interface because these binding sites are subjected to natural selective pressures to maintain the interaction over evolutionary

time<sup>36</sup>. Multiple sequence alignment (MSA) and phylogenetic tree based methods have been used by several groups to study conservation patterns at interfaces<sup>10, 27-30, 37</sup>. While Lichtarge and co-workers have looked for patterns in functional groups of homologous sequences of a protein family<sup>25</sup>, Lockless & Ranganathan have shown energetic coupling of residues at the interface with distant residues and that these coupling pathways were evolutionarily conserved<sup>38</sup>. These methods were, however, based on the idea that residues at the interface are more conserved than the rest of the protein<sup>38</sup>. Grishin & Phillips tested this idea and showed that active site residues were highly conserved and evolved at a slower than average rate<sup>34</sup>. Interfacial and core residues on the other hand were not as well conserved and evolved as rapidly as the rest of the protein surface. Interfacial residues, although less conserved than active sites, were still more conserved than the surface residues<sup>34</sup>. Buyong et al, in their correlation studies, show that energy hot spots are generally conserved and that these residues can help distinguish between interfacial and surface residues<sup>39</sup>. Groups employing sequence conservation for interface detection have had mixed success depending on precisely which features they employ. To date the EPPIC server of Duarte appears to achieve one of the highest successes at an impressive 89% accuracy on the set of Ponstingl<sup>10, 40</sup>; however they, like others, combine several hybrid approaches with sequence conservation.

In previous work<sup>41</sup>, we operationally defined Functionally Linked Interfaces of Protein (FLIP) structures as interfaces known to be critical to biological function. This would include interfaces known to oligomerize *in vivo* or ones in which mutation, chemical modification, or other physical change to the PPI alters the native biological

function. Similarly, PPIs with no such identified alteration in function, such as known obligate monomers, crystal contacts, or artificial constructs, were defined as Functionally uncorrelated Contacts (FunC). In our prior work, by combining geometric relationships from atomic resolution structures and energetic data from computational alanine scanning ( $\Delta\Delta G_{\text{ala}}$ ), we could distinguish FLIPs from FunCs with 76-80% accuracy<sup>41</sup>. In that work, while both FLIPs and FunCs demonstrated strong and weak energetic values, overall FLIPs demonstrated much stronger energies than FunCs. In addition, these energies were organized radially around the geometric centers of the interface. FunCs, on the other hand, tended to have smaller, energetically weaker, and less organized features. The presence of a central tendency, which we term the Energy Centrality Relationship or ECR, in FLIPs suggests a progressive re-organization relative to the interface center, possibly through selective pressure on interface maintenance. This ECR concept is supported by the fact that radial symmetry of sequence variation in functionally important regions is also observed in Evolutionary Traces from Lichtarge and colleagues<sup>25</sup>. It is also in qualitative agreement with work by Janin<sup>8, 23</sup> and Capitani<sup>10</sup> showing interface rims are less important for PPI detection than well packed and buried cores<sup>17</sup>.

FLIP interfaces in general exhibited more specificity and an overall greater central tendency than FunCs, suggesting that FLIPs experience selective pressure to maintain a stable interface. Over evolutionary time these selective pressures might generate a radially symmetrical pattern, with more conserved residues at the center of the interface in FLIPs. FunCs on the other hand experience limited or no selective pressure and should have more random distribution of conserved residues.

To test this concept, we analyzed the FLIPdb set of PPI from our earlier work. For each interface, we used Lichtarge's Evolutionary Trace method<sup>37</sup> and our own geometry calculations to determine the residue conservations patterns in an interface as a function of geometric distribution in the interface.  $\rho$ -score and # states in multiple sequence alignment values were regressed against distance from the center of the interface. Principal component analysis<sup>42</sup> and K-means clustering<sup>43</sup> were used to identify 5 spatial characteristics of sequence conservation that could distinguish FLIPs from FunC, with 69% accuracy. The same characteristics were used to analyze a test set of 18 sequence dissimilar interfaces, though only with 61% accuracy. To test for compositional bias, random sub-sample analysis was used and indicated little bias. Overall, as with energetic criteria, sequence conservation also could distinguish FLIPs from FunC, and protein interfaces with similar biological functions were found to group together. FLIPs were indeed found to have more stable cores and more variable edges. In head-to-head and hybrid comparisons, sequence conservation was found to be not as strong a discriminator as previous energetic based features.

Though slightly lower in accuracy, this work is useful, particularly in comparison to our prior work as we show that while having a central tendency and showing more conservation at the core than at the edges, FLIPs nevertheless had more variability than would otherwise be expected, perhaps demonstrating conservation of energy at the expense of sequence stability.

## Results

### *Database composition – FLIPdb-ET*

Here, we analyzed the sequence conservation patterns of the interfaces in our<sup>44</sup> interface database, FLIPdb. The protein interfaces studied were the same as in our previous work, although 6 interfaces generated pathological errors and were excluded (Table 3.1). In 4 proteins (1daz\_AD (inhibitor), 1aw7\_CD (FunC), 1pky\_AB, and 1pky\_BD (Enzyme)), the Evolutionary Trace (ET) method<sup>25, 37, 44</sup> failed due to a lack of sufficiently similar sequences being generated by BLAST<sup>45</sup>. In the remaining 2 proteins (1tzi\_AV, 1tzi\_BV (Ab-Ag)), the ET method identified sequences whose conservation scoring patterns were more than twice the upper confidence limit during Mahalanobis outlier analysis at an  $\alpha = 0.05$  and were therefore removed from further consideration. The version of FLIPdb used for the analysis of ET data (FLIPdb-ET), therefore, consists of 154 interfaces (Table 3.1). Since the rejected interfaces were from different functional sub-categories and comprised only 4% of the original data, we did not expect their exclusion to have substantial impact on data analysis and conclusions drawn from it. The test set used to validate was the same as in our previous work<sup>41</sup>.

### *$\rho$ distribution in PPI*

We used Lichtarge's Evolutionary Trace method<sup>25</sup> to determine two measures of sequence conservation, the degree of amino acid variation at a given multi-sequence alignment position ( $\rho$ ) and the number of amino acid states for a residue in a multiple sequence alignment (states). Simplistically, the  $\rho$  score is a Shannon entropy hybridized with a tree branch ranking trace<sup>37</sup>. These sequence conservation metrics were also

compared to their geometric distribution in the interface ( $r$ , distance ( $\text{\AA}$ ) of each residues'  $C\alpha$  from the mean  $C\alpha$  position). The distributions of the two metrics were compared between FLIPs and FunCs. Two representative examples are shown in Figure 3.1, in which we compared analysis of the structure of a FunC (PDBid: x1brw1, chains A and Z – a crystal contact of the PDB structure 1brw – a nucleoside pyrolase) and a FLIP (PDBid: 1osj, chains A and B – an oxidoreductase dimer. Histogrammed contours of the variation in  $p$  are plotted on major axes of the interface geometry (as determined by Principal Component Analysis) (Figures 3.1A and B). These analyses indicate that, in the FLIPs, conserved residues are more centrally located and exhibit a radial distribution pattern (Figure 3.1B). FunCs, on the other hand, have more random distribution of conserved residues and generally do not display a radial distribution pattern or central tendency (Figure 3.1A). Conservation scores and distances for interfaces in the FLIPdb were then fit to first order polynomial line via linear regression, generating slopes and intercepts. The distribution of the upper and lower quartiles of slopes are shown in Figure 3.1C. FLIPs tend to populate the positive slope region (mean 0.08, range -2.63:1.09) while FunC, (mean -0.109, range -1.82:0.94) tended to populate the negative slope region. One-way ANOVA at an  $\alpha = 0.10$  and using Kruskal-Wallis tests show slopes of the # of states are significantly different ( $P \leq 0.029$ ) between FLIPs and FunCs. Positive slopes are indicative of a central tendency with less sequence variation at the interface center and more sequence variation at the periphery. The average negative slope of FunCs suggest that they have just the opposite behavior with more sequence variation at

the interface centers, in agreement with<sup>8, 10, 17, 23</sup>. Interestingly, Ab-Ag interactions had amongst the most negative slopes (mean -0.772, range -2.63:0.217) and were distinctly more similar to FunCs than to FLIPs. Summary statistics are described in the next section.

### Evolutionary and Geometric Features

We analyzed sequence variation (number of states and  $\rho$ ) and its correlation with distance from Center of Interface (CoI) to represent PPIs. By regression of these variations to geometry, we generated 4 features: the sequence variation as a function of distance (slope\_state and slope\_ $\rho$ ), and the sensitivity to variation at the interface center (intcpt\_state and intcpt\_ $\rho$ ).

Three (3) features describe net sensitivity to variation at the interface: collective variation of all residues at the interface (sum\_states, sum\_ $\rho$ ), mean variation for all residues at the interface (avg\_states, avg\_ $\rho$ ), and total number of residues at the interface (#total) (Figure 3.2). Non-parametric one-way ANOVA at an  $\alpha = 0.10$  with Kruskal-Wallis tests show all features based on # states were significantly different between FLIP and FunC ( $P_{\text{slope}} \leq 0.02$ ,  $P_{\text{intcpt}} \leq 0.004$ ,  $P_{\text{sum}} \leq 0.001$ ,  $P_{\text{avg}} \leq 0.01$  and  $P_{\text{total}} \leq 0.0001$ ). Somewhat unexpectedly, of the features based on  $\rho$ -scores from Evolutionary Trace, only the intercept and averages were significantly different ( $P_{\text{intcpt}} \leq 0.0006$ ,  $P_{\text{avg}} \leq 0.003$ ). Though statistically distinct, individually, none of these features were found to sufficiently correlate with FLIP or FunC designations such that a single feature could be used to identify the category.

### ***Principal component analysis (PCA)***

With no individual feature easily discriminating FLIP from FunC, a multi-factorial approach of Principal Component Analysis of the 9 features for all 154 PPI in the training set was undertaken. Features with lower eigenvector coefficients were systematically removed until the first two principal components (PCs) reproduced 80% of the normalized data variations (Figures 3.2 and 3.3). Terms based simply on the number of amino acid states, rather than  $\rho$ , demonstrated larger eigenvector coefficients (i.e. better correlations). While this was consistent with the ANOVA, this still surprised us, as traces and Shannon entropies generally produce more useful and reliable comparisons than simple residue variation. One source of the stronger correlations with number of states may have been the fact that  $\rho$  is derived from the branches and log frequencies of one chain whereas our geometric analyses essentially mixed the  $\rho$  of two chains. Simple variation in states is less dependent on within chain clustering and may have been more resilient to our simplistic analyses. In the future work, a multi-point coupled analysis like that of Ranganathan<sup>38</sup> may improve the relationship with  $\rho$ . Because of this poor correlation and lack of statistical difference,  $\rho$ -based terms were eliminated. The resulting PCA reproduced 84.1% of the variance in the first two principal components.

### ***K-means clustering***

To ascertain if the features identified could identify groups of FLIPs and FunCs, we performed a K-means cluster analysis. The K-means method clusters data into  $k$  clusters based on specific criteria (features identified from loading plot). The clustering analysis with a two-cluster assumption ( $k = 2$ ) of our data produced two clusters whose

centers straddled the origin for both PC1 and PC2, indicating opposing correlation trends (Figure 3.3). Cluster 1 (hereafter called the FLIP cluster as it contained more FLIPs than FunCs) contained 92% and 69% of FLIP and FunC interfaces, respectively. Cluster 2 (hereafter the FunC cluster as it contained more FunCs than FLIPs) contained 8% of all FLIP and 30% of FunC interfaces. Further inspection indicated that PPIs in the FLIP cluster were more positive on PC1. Limited trends emerged from our analysis, including the fact that distinct sub-category groupings of FLIPs and FunCs were observed and that Ab-Ag and inhibitors cluster with FunCs, as was seen in energetic analysis. A second clustering of only the 26 interfaces in the FunC cluster was performed to further analyze these initial trends. The same set of variables was used in the second clustering and it produced PCs that represented 87% of the data variation in the first two PCs (Figure 3.2 C). In both clusterings, FunC, XFunC, and Ab-Ag sub-categories dominated the FunC cluster and the PCs produced had similar eigenvector correlation coefficients for these sub-categories. From these results we understood that interfaces in the two clusterings had similar data dependencies. K-means clustering of the second PCA also produced two clusters; cluster 1 contained 50% of the unassigned FLIPs and 3% of FunC interfaces again demonstrating discrimination. Over two-rounds of clustering, 95% of the FLIPs positively correlated with the chosen features and 28% of the FunCs negatively correlated with the features. FLIPs were defined by larger interfaces that correlated with lower average #states (more conservation) at the interface.

### ***Accuracy and Matthews correlation***

Analysis of the two rounds of PCA of the training set PPI data achieved accuracies (propensity to correctly identify FLIP or FunC) of 68% and 77% in each round of clustering. Over both rounds the total accuracy was 69% (Table 3.3). This accuracy was rather lower than might be expected based both on the success of our earlier energetic analysis (76-80%) and on the success of other sequence conservation methods that achieved accuracies as high as 84-89%<sup>9, 10</sup>. However, the Matthews Correlation Coefficient (MCC), a measure of how well a binary classification matches the data, was 0.29 in clustering round one and 0.54 in clustering round 2, and 0.30 overall. Again, this was slightly lower than using energetic criteria but still consistent with a two-cluster approximation.

### ***Cross-validation testing***

While analysis of the training set favorably predicted distinct feature set correlations between FLIPs and FunCs, it was possible that the relationship was training set dependent. To test this, we repeated the analyses on the same test set of 18 unrelated proteins as used in our previous work<sup>41</sup>. This set of 18 interfaces was design to have less than 70% sequence conservation with the training set. No new PCA or K-means clustering analyses were undertaken; rather the features of the test set were projected through the PCs of the training set. Test set projections are shown in Figures 3.3C and D. The overall accuracy was lower than the training set at 61% (Table 3.3). Similarly, MCC values of the test set were much lower than the training set at -0.22 (Table 3.3). In previous work based on energetic analysis, this test set gave us superior results (78%) and

extremely consistent results with the training set. This test set was small in number and was slightly enriched in Ab-Ag interactions, both factors potentially reducing our success rates.

Because of this possible training set dependence, we evaluated the compositional bias of our training set for sequence variation using random sub-sampling validation (Table 3.5). Sub-samples of the training set were generated randomly in triplicate for subsets ranging from 90% down to 20%. Regression analysis at an  $\alpha = 0.10$  for 1<sup>st</sup> through 6<sup>th</sup> order polynomial fits of number of PPI vs. Accuracy show substantial *Lack of Fit* error and a lack of statistical significance for all but the 4<sup>th</sup> order fit, which has a tepid  $P \leq 0.5$ . Overall, from regression analysis we can infer that little compositional bias exists until the number of PPI falls substantially below 50% of the training set. As our test set had only 18 interfaces (12% the size of our training set) and several of those were Ab-Ag, we may have fallen into this limit. In effect, our test may have been biased. This was unexpected as the test set was highly consistent with training results in previous energetic analysis. These results suggest that conservation data is potentially more training set dependent than energetic data and may be less suitable for FLIP/FunC discrimination.

### ***Energetic and evolutionary features***

Our previous work using energy distributions from computational alanine scans in PPI had yielded seven features that can distinguish FLIPs from FunCs with 76-80% accuracy<sup>41</sup>. Even though analysis of features obtained from our calculations of sequence conservation and geometric data seemed to be more training set dependent than energetic analysis, we explored combining features obtained from energy and geometric

calculations with the sequence conservation features. No p values were tested as they had previously been found to be less diagnostic than #states. The 11 combined features accounted for 64% of the total variance in the first two PCs (Figure 3.4A). Three features, *intcpt\_states*, *avg\_states* and *slope\_states*, were excluded from further analysis due to limited contributions (coefficients less than 0.6) to the overall data variation compared to energetic features. Though these values were distinct in conservation analysis alone, they were not as strong as energetic values in describing the data. The net effect was adding only the *sum\_states* feature to the energetic features. A loading plot of these features shows PCs 1 and 2 accounted for 85% of the variance (Figure 3.4B).

The results of clusterings of these 154 interfaces using both 11 features and 8 features with  $k = 2$ , are shown in Table 3.4. All 11 features produced clusters with 75% accuracy and essentially the same discriminatory characteristics as energy alone (Table 3.4.A). However, when energy features were coupled with the one single strongest sequence variation trait (*sum\_states*, Table 3.4.B), the accuracies essentially dropped below sequence conservation accuracies almost to random guessing. Oddly, the MCC's remained high, continuing to suggest a two-category model was appropriate. By themselves, each feature appears capable of discriminating FLIPs from FunCs. However, these data suggest that hybridizing energetic and sequence conservation methods cancel the sensitivity of the other. Overall this suggests that energetic-based criteria are more useful for interface detection than sequence variation.

## **Discussion**

### ***Evolutionary trace can distinguish FLIP from FunC interfaces***

Our comparison of evolutionary data with interface geometries during analysis of the training set seems to suggest FLIP interfaces are likely to have different residue characteristics than FunC interfaces. FLIP PPIs could be distinguished from FunC PPIs with 69% accuracy (Table 3.2). This result is also supported by the fact that PPIs of the same functional sub-category generally have similar PC projection values such that they cluster (Figure 3.3), even though FLIP and FunC categories had substantial overlap in characteristics. Our cross-validation methods provided mixed messages. Random sub-sampling had little compositional bias until random compositions of less than 50% of the training set were tested, while the test set of 18 PPI with less than 70% sequence identity achieved limited success with sequence traits but high success with energetic traits. The relationship between sequence conservation and geometry, as we have used it here, does not seem to be as reliable a FLIP/FunC discriminator as energetic criteria. Irrespective of that, certain overall patterns did seem to consistently emerge from our analyses from both training and testing sets (Figures 3.2, 3.3).

### ***Physical interpretation***

Consistent with our hypothesis of FLIPs having a central tendency, FLIPs had positive correlation with the regression slope of number of states, meaning that less variation occurs near the interface center relative to the edges. FunCs on the other hand seemed to be much more variable in conservation patterns and indeed were generally

negatively correlated with slope, meaning they have more variation towards the center. This is also consistent with the patterns observed by Lichtarge and colleagues in numerous systems<sup>37, 44, 46</sup>.

Contrary to our expectations, our hybridization of energetic and sequence conservation studies was actually worse than using energy or sequence traits by themselves (Figure 3.4A&B). We suggest this cancelling out of sensitivity in FLIP interfacial residues may be due to the co-evolutionary “optimization” of the interface. During selection at an interface, different co-evolved residue interactions can equally well improve energetic stability. This would have the effect of increasing sequence variation and disrupting correlation with geometric patterns. This type of mixed population should be amenable to detection by Evolutionary Trace methods if the correlation between chains can be established.

### ***Implications***

The geometric distribution of conservation features can distinguish between FLIPs and FunC but only with limited accuracy. While several conservation trends consistent with the findings of others emerge from our analysis, the discriminatory power is less than that of energetic criteria. Thus, sequence conservation methods, as we implement them here, may be less reliable at FLIP/FunC discrimination. This mixed success using conservation methods is consistent with prior findings<sup>38, 39, 47</sup>. Sequence conservation is not simply a result of chemistry but also genetic control mechanisms that may not be represented by positional variation within the protein.

In our previous ECR work, we saw enzymes and Ab-HL to be strongly associated with FLIPs; Ab-Ag, though classified FLIP, strongly grouped with FunCs, and Structurals clustered with both FLIPs and FunCs. Cluster analysis on conservation data also generated results consistent with our ECR work<sup>41</sup>.

Clustering shows that our operational definition of FLIP/FunC may not be appropriate for all functional sub-categories. Ab-Ag, for example, tends to cluster with the FunCs. These findings are consistent with our previous energy-based studies. We suggest that sequence variation alone, as a function of geometric distribution at the interface, is less likely to cleanly differentiate FLIP from FunC and that energetic criteria be used whenever possible.

Additional studies excluding Ab-Ag interfaces were performed to assess if these interfaces improved or obscured data discrimination. In this modified clustering analysis we achieved an accuracy of 75% and a MCC of 0.32. These were promising results as they not only produced a larger success of predicting FLIPs and FunCs but also clarified the nature of these interfaces. It could mean that though Ab-Ag interactions have a selection component during somatic cell hypermutation and B-cell selection, it may be a different evolutionary-like process than that experienced by enzymes.

## **Conclusions**

Here, we show that protein-protein interfaces associated with biological functions maintain distinct physical features from contacts uncorrelated with function. As in previous work, here we show that FLIPs demonstrate a strong central tendency with more conserved central cores that have stronger binding and edges that are much more variable

and contribute less to interface stability. We show that while having a central tendency and showing more conservation at the core than at the edges, FLIPs nevertheless had more variability than would otherwise be expected, perhaps demonstrating conservation of energy at the expense of sequence stability. We find that energetic-based methods are more diagnostic than similar sequence-based approaches.

## **Methods**

### ***FLIPdb-ET dataset***

FLIPdb was constructed using atomic resolution structural data available in PDB<sup>48</sup> and additional information available in literature and structural information databases to categorize them into functionally-linked interfaces (FLIP) and functionally uncorrelated interfaces (FunC). FLIPdb consists of 160 PPIs and the version of FLIPdb used for this project consists of 154 interfaces analysis issues with six interfaces. The interfaces omitted were 1tzi\_ab, 1tzi\_av (Ag-Ab), 1daz\_ad (Inhibitor), 1aw7\_cd (FunC), and x1brw1 (XFunC) (Table 3.1). For details on construction of FLIPdb refer to<sup>41</sup>.

### ***Evolutionary trace***

The Evolutionary Trace (ET) method of Lichtarge et al<sup>25</sup>, used to evaluate all interfaces in FLIPdb-ET, is a rapid method for ranking the evolutionary importance of residues. Simplistically, the p score of ET is a Shannon entropy hybridized with a tree branch ranking trace. All ET analysis was performed using Report\_Maker at the ET server<sup>49</sup>.

### ***Interfacial geometry***

Kortemme and Baker's<sup>50</sup> definition of interface was used in our analysis. In brief, it is residues with atoms within 4 Å of partner chain atoms or residues with C $\beta$  buried during interface formation. The geometric distribution of residues at the interface was studied by calculating the distance ( $\Delta r$ , Å) between residue C $\alpha$  at the interface and the CoI. The CoI is defined as the mean of all interfacial C $\alpha$  positions.

### ***Geometry analysis***

Linear regressions of  $\rho$  or #states values with  $\Delta r$  were calculated using GNU PLOT<sup>51</sup> and our own software. The calculations provided 7 features for each interface: the slope (slope\_ $\rho$  and slope\_states), intercept (intcpt\_ $\rho$  and intcpt\_states), net variation of all residues at the interface (sum\_ $\rho$  or sum\_state), mean variation of all residues at the interface (avg\_ $\rho$ , avg\_states), and total number of residues in the interface (#total).

### ***Principle component analysis (PCA)***

Principle component analysis of the variation of  $\rho$ , variability and geometric feature data for all PPI was undertaken using JMP<sup>52</sup>. PCA determines a set of linearly-uncoupled eigenvectors from normalized correlations between variables that progressively describe the largest sources of variance in a data set<sup>42</sup>. The eigenvector coefficients for each principal component vector indicate the relative correlation between each feature and the overall variation of all features. In this work, we sought to identify the set of features that would describe more than 80% of the total set variation in the first

two principal components (PCs) such that we could use these PCs to discriminate between FunC and FLIP data. The results from these PCA analyses are shown in Figure 3.2 and Table 3.1.

### ***K-means clustering***

K-means clustering<sup>43</sup> is a data analysis method that clusters observations into a specific number of clusters by attempting to find the point(s) that have the lowest mean variation from the other input data. When combined with PCA, the combination of features that allows input data to be clustered can be identified. In this work, two clusters were specified and the correlations between cluster and functional category determined (Figure 3A,B, Table 3.3).

### ***Accuracy and Matthews correlation coefficient***

The following measures were used to assess the performance of our clustering analysis: *Accuracy* (ACC), the propensity to correctly identify FLIP or FunC:

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

and *Matthews correlation coefficient* (MCC), a measure of how much a set of predictive data agrees with a two-state model:

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

where,

TP = correctly predicted as FLIPs, TN= correctly predicted as FunCs

FP = wrongly predicted as FLIPs, FN = wrongly predicted as FunCs.

These values are shown in Table 3.3.

### **Acknowledgements**

We thank Texas Woman's University Office of Technology for high performance computing resources. The work of SS on this project was supported by internal funding from TWU (Research Enhancement Program, Multidisciplinary Research Program, and Art & Sciences Research Development Fund).

## References

1. Phizicky EM, Fields S (1995) Protein-protein interactions: methods for detection and analysis. *Microbiol Rev* 59:94-123.
2. Wells CT (1995) A hot spot of binding energy in a hormone-receptor interface. *Science* 267:383-386.
3. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, et al (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403:623-627.
4. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams S, Millar A, Taylor P, Bennett K, Boutilier K, et al (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415:180-183.
5. Gavin A, Bösch M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon A, Cruciat C, et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415:141-147.
6. Bahadur RP, Zacharias M (2008) The interface of protein-protein complexes: analysis of contacts and prediction of interactions. *Cell Mol Life Sci* 65:1059-1072.
7. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2003) Dissecting subunit interfaces in homodimeric proteins. *Proteins* 53:708-719.
8. Dey S, Pal A, Chakrabarti P, Janin J (2010) The subunit interfaces of weakly associated homodimeric proteins. *J Mol Biol* 398:146-160.
9. Zhu H, Domingues FS, Sommer I, Lengauer T (2006) NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics* 7:27-27.
10. Duarte JM, Srebniak A, Schärer M, A., Capitani G (2012) Protein interface classification by evolutionary analysis. *BMC Bioinformatics* 13:334-334.
11. Krissinel E, Henrick K Detection of Protein Assemblies in Crystals In: R. Berthold M, Glen R, Diederichs K, Kohlbacher O, Fischer I, Eds. (2005) , Springer Berlin Heidelberg, pp 163-174.
12. Valdar WS, Thornton JM (2001) Conservation helps to identify biologically relevant crystal contacts. *J Mol Biol* 313:399-416.

13. Jones S, Thornton JM (1996) Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* 93:13-20.
14. Mintseris J, Weng Z (2005) Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A* 102:10930-10935.
15. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. *J Mol Biol* 285:2177-2198.
16. Keskin O, Gursoy A, Ma B, Nussinov R (2008) Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chem Rev* 108:1225-1244.
17. Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. *J Mol Biol* 280:1-9.
18. Sonavane S, Chakrabarti P (2008) Cavities and atomic packing in protein structures and interfaces. *PLoS Comput Biol* 4:e1000188-e1000188.
19. Malod-Dognin N, Bansal A, Cazals F (2012) Characterizing the morphology of protein binding patches. *Proteins* 80:2652-2665.
20. Swapna LS, Bhaskara RM, Sharma J, Srinivasan N (2012) Roles of residues in the interface of transient protein-protein complexes before complexation. *Sci Rep* 2:334-334.
21. Tuncbag N, Kar G, Keskin O, Gursoy A, Nussinov R (2009) A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief Bioinform* 10:217-232.
22. Verdonk ML, Giangreco I, Hall RJ, Korb O, Mortenson PN, Murray CW (2011) Docking performance of fragments and druglike compounds. *J Med Chem* 54:5422-5431.
23. Chakrabarti P, Janin J (2002) Dissecting protein-protein recognition sites. *Proteins* 47:334-343.
24. Valdar WS, Thornton JM (2001) Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* 42:108-124.
25. Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 257:342-358.

26. Valdar WS (2002) Scoring residue conservation. *Proteins* 48:227-241.
27. Scharer MA, Grutter M, Capitani G (2010) CRK: an evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts. *Proteins JID* - 8700181.
28. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2004) A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol* 336:943-955.
29. Elcock AH, McCammon JA (2001) Identification of protein oligomerization states by analysis of interface conservation. *Proc Natl Acad Sci U S A* 98:2990-2994.
30. Guharoy M, Chakrabarti P (2005) Conservation and relative importance of residues across protein-protein interfaces. *Proc Natl Acad Sci U S A* 102:15447-15452.
31. Valencia A, Pazos F (2002) Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol* 12:368-373.
32. Levy ED (2007) PiQSi: protein quaternary structure investigation. *Structure* 15:1364-1367.
33. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci* 13:190-202.
34. Grishin NV, Phillips MA (1994) The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Sci* 3:2455-2458.
35. Bordner AJ, Ab-Agyan R (2005) Statistical analysis and prediction of protein-protein interfaces. *Proteins* 60:353-366.
36. Guharoy M, Chakrabarti P (2010) Conserved residue clusters at protein-protein interfaces and their use in binding site identification. *BMC Bioinformatics* 11:286-286.
37. Mihalek I, Reš I, Lichtarge O (2004) A Family of Evolution–Entropy Hybrid Methods for Ranking Protein Residues by Importance. *J Mol Biol* 336:1265.
38. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286:295-299.

39. Ma B, Elkayam T, Wolfson H, Nussinov R (2003) Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A* 100:5772-5777.
40. Henrick PH (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins* 41:47-57.
41. Sudarshan S, Kodathala SB, Mahadik AC, Mehta I, Beck BW (2014) Protein-protein interface detection using the energy centrality relationship (ECR) characteristic of proteins. *PLoS One* 9:e97115-e97115.
42. Pearson K (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* 2 11:559-572.
43. Hartigan JA (1973) Clustering. *Annu Rev Biophys Bioeng* 2:81-101.
44. Mihalek I, Res I, Lichtarge O (2006) Evolutionary trace report\_maker: a new type of service for comparative analysis of proteins. *Bioinformatics* 22:1656-1657.
45. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403-410.
46. Morgan DH, Kristensen DM, Mittelman D, Lichtarge O (2006) ET viewer: an application for predicting and visualizing functional sites in protein structures. *Bioinformatics* 22:2049-2050.
47. Grishin NV, Phillips MA, Goldsmith EJ (1995) Modeling of the spatial structure of eukaryotic ornithine decarboxylases. *Protein Sci* 4:1291-1304.
48. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235-242.
49. <http://mammoth.bcm.tmc.edu>.
50. Kim KT (2004) Computational alanine scanning of protein-protein interfaces. *Sci STKE* 2004:pl2-pl2.
51. Williams T, Kelley C (2011) .
52. SAS Institute Inc., Cary, NC (1989-2012) JMP. 10.

**Table 3.1. FLIPdb-ET: Protein Interfaces and Functional Categories**

Category	Sub-category	Protein Interfaces
<b>FLIP</b>	<b>Ab-Ag</b>	1acy_HP 1acy_LP 1adq_AH 1adq_AL 1tzi_AV 1tzi_BV 2bse_AE 2bse_AF 2bse_BD 2bse_BE 2bse_CD 2bse_CF
	<b>AbHL</b>	1acy_HL 1adq_HL 1gaf_HL 1tzi_AB 1wtl_AB
	<b>Enzyme</b>	1b5e_AB 1biq_AB 1bis_AB 1bjw_AB 1bkb_AB 1bmd_AB 1brw_AB 1bsl_AB 1bsr_AB 1cg2_AD 1cg2_BC 1cnz_AB 1coz_AB 1daz_CD 1hjr_AC 1hjr_BD 1itv_AB 1ivy_AB 1nhk_RL 1nsy_AB 1oro_AB 1osj_AB 1pgt_AB 1pky_AB 1pky_AC 1pky_BD 1pky_CD 1pre_AB 1qks_AB 1qr2_AB 1r2f_AB 1scu_AB 1scu_DE 1sft_AB 1slt_AB 1smn_AB 1sox_AB 1ubs_AB 1vfr_AB x1ubs1_BZ
	<b>Inhibitor</b>	1cmi_AC 1cmi_BD 1daz_CB 1daz_DA 1ppf_EI
	<b>Receptor</b>	1cdc_AB 1eaj_AB 1fcc_AC 1h0t_AB 1nrv_AB 2bse_AB 2bse_AC 2bse_BC 2cii_AB 2cii_AC
	<b>Regulator</b>	1awi_AB 1awi_AP 1awi_BP 1edh_AB 1f13_AB 1f4v_AD 1psr_AB 1qfh_AB 1tx4_AB 2arc_AB 2trc_BG 2trc_BP
	<b>Structural</b>	1cmb_AB 1f95_AB 1f95_AC 1f95_BD 1sa0_AB 1sa0_AE 1sa0_BE 1sa0_CD 1sa0_CE 1sa0_DE 1tub_AB 2bkh_AB 2bki_AB 3kin_AB 3kin_BD 3kin_CD
<b>FunC</b>	<b>FunC</b>	1atn_AD 1aw7_CD 1azt_AB 1bin_AB 1bkz_AB 1bo5_OZ 1c02_AB 1cmi_AB 1cqx_AB 1f4v_AB 1gqp_AB 1hjr_AB 1lcu_AB 1m6k_AB 1naw_AB 1nmt_AB 1nmt_BC 1oio_AB 1ome_AB 1scu_AE 1scu_BD 1scu_BE 1vio_AB 1xca_AB 3k3e_AB
	<b>XFunC</b>	x1b8e4_AE x1brw1_AZ x1brw3_BY x1bsf8_AJ x1bsr1_AB x1dto6_AG x1dv82_AC x1e872_AC x1erv2_AC x1gaf1_HZ x1gaf2_LY x1gaf3_HL x1ihk3_AD x1neu4_AE x1oal6_AG x1oro1_AZ x1oro3_AZ x1ppf1_EI x1ppf2_EZ x1ppf3_EI x1py94_AE x1sox2_BY x1tx41_AB x1tx42_AZ x1tx43_AB x1uq54_AE x1vfr2_AZ x1vfr3_BY x1wtl1_BZ x1wtl2_BZ x1xca1_AZ x1xca2_BY x1xca3_AB x3bri2_AC x3il85_AF

\*Light grey interfaces were rejected from the analysis due to lack of sufficiently similar sequences.

**Table 3.2.** *Summary of Protein and Protein Interface Counts in FLIPdb-ET*

Function		Training Set			Test 18 Set		
Category	Sub-categories	PDB Structures	Protein chains	Interfaces	PDB Structures	Protein chains	Interfaces
FLIP	Ab-Ag*	4	14	10	1	6	5
	AbHL*	5	10	5	1	4	2
	Enzyme	33	73	38	2	4	2
	Structural	7	21	16	1	2	1
	Receptor	7	16	10	1	2	1
	Regulator	9	20	12	1	2	1
	Inhibitor	3	9	4	1	2	1
	<b>Total</b>	63	150	95	7	18	13
FunC	FunC	21	45	24	-	-	-
	XFunC‡	23	44	35	5	10	5
	<b>Total</b>	43	87	59	5	10	5
<b>Total</b>		93	214	154	7	19	18

\* Protein chains are common to multiple sub-categories though interfaces are distinct.

‡ Interfaces are constructed from existing FLIPs through coordinate transformations arising from the symmetry of the source X-ray crystal structure (XFunCs).

FLIPdb-ET contains 154 interfaces in 93 structures involving 214 individual proteins chains. These interfaces have been assigned to FLIP or FunC functional categories and 9 functional sub-categories based on a review of the literature<sup>41</sup>. Due to the reuse of some chains, the totals represented in the first two columns do not sum across sub-categories.

**Table 3.3.** *Accuracy of Clustering in Training and Testing Sets*

	<b>True Positive (TP) †</b>	<b>False Positive (FP) †</b>	<b>False Negative (FN) †</b>	<b>True Negative (TN) †</b>	<b>Accuracy</b>	<b>MCC</b>
<b>Training Set</b>	1 <sup>st</sup> Clustering					
	87	41	8	18	68.18	0.29
	2 <sup>nd</sup> Clustering					
	4	2	6	16	76.92	0.54
<b>Total</b>	91	43	6	16	69.48	0.30
<b>Testing Set</b>	1 <sup>st</sup> Clustering					
	9	4	4	1	55.56	-0.11
	2 <sup>nd</sup> Clustering					
	2	1	2	0	40.00	-0.41
<b>Total</b>	11	4	2	1	61.11	-0.22

† TP: FLIP correctly predicted

† TN: FUNC correctly predicted

† FP: FUNC identified as FLIP

† FN: FLIP identified as FUNC

\*Total FNs and TNs do not add sum across 1<sup>st</sup> and 2<sup>nd</sup> clustering. TP and FP sum, but FN and TN arise only from 2<sup>nd</sup> clustering

The accuracy and Matthews correlation coefficient (MCC, a measure of the quality of a binary classification) of the results of the clusterings shown in Figure 3.3 are indicated. Analysis of two rounds of K-means clustering of the training set data indicated that the overall accuracy of predicting FLIP is 68% and 61% in both testing and training sets, respectively. TP are more readily identified in both Testing and Training sets (68% and 73% *sensitivity*, respectively). 92% of the total FLIP population was identified in the first round of clustering. TN are less well identified (28% and 25% *negative predictive values*, respectively). Our two-category approach had an MCC of 0.30 for the training set and an MCC of -0.41 for the test set

**Table 3.4.** *Accuracy of Clustering upon Combining ECR with ET.*

<b>A</b>	<b>True Positive (TP) †</b>	<b>False Positive (FP) †</b>	<b>False Negative (FN) †</b>	<b>True Negative (TN) †</b>	<b>Accuracy</b>	<b>MCC</b>
<b>ECR + ET All 11 Features</b>	1 <sup>st</sup> Clustering					
	57	5	38	53	72.07	0.51
	2 <sup>nd</sup> Clustering					
	23	18	15	36	64.13	0.29
<b>Total</b>	80	23	15	36	75.32	0.46
<b>B</b>  <b>ECR + ET 8 best Features</b>	1 <sup>st</sup> Clustering					
	7	1	88	58	42.20	0.12
	2 <sup>nd</sup> Clustering					
	39	0	53	54	63.69	0.46
<b>Total</b>	46	1	53	53	64.93	0.47

† TP: FLIP correctly predicted

† TN: FUNC correctly predicted

† FP: FUNC identified as FLIP

† FN: FLIP identified as FUNC

\*Total FNs and TNs do not add sum across 1<sup>st</sup> and 2<sup>nd</sup> clustering. TP and FP sum, but FN and TN arise only from 2<sup>nd</sup> clustering

The accuracy and MCC of the clustering results of combining Energetic and Sequence features (Figure 3.4) are indicated. Results are shown for (A) the complete overlap of ECR and ET (11 features) and (B) for the combination of only the most correlating features (essentially ECR plus sum\_states from this work). The overall accuracy of predicting FLIPs in the training set is 75% and 65% for each set of features, respectively. (A) FLIPs continue to be readily identified in both approaches (78% and 98% *sensitivity*, respectively). (B) Interestingly combination of the most maximally correlating features actually reduced the accuracy with respect to BOTH methods alone. Seemingly the reverse, MCC's of 0.46 and 0.47 continue to suggest that a two-category approach is representative for the current training set.

**Table 3.5: Random Sub-sample Validation of FLIPdb-ET Training Set**

<b>% FLIPdb (# of Interfaces)</b>	<b>Trial</b>	<b>TP†</b>	<b>FP†</b>	<b>FN†</b>	<b>TN†</b>	<b>Accuracy</b>	<b>MCC</b>
<b>100</b>	1	91	41	6	16	69%	0.30
<b>90 (138)</b>	1	85	32	6	15	72%	0.33
	2	83	38	5	12	68%	0.26
	3	80	38	5	15	68%	0.30
<b>80 (123)</b>	1	70	27	5	21	73%	0.44
	2	71	35	6	12	66%	0.24
	3	65	27	10	20	69%	0.33
<b>70 (108)</b>	1	30	20	36	22	48%	-0.02
	2	33	25	31	19	48%	-0.05
	3	80	29	6	12	72%	0.29
<b>60 (92)</b>	1	53	19	5	15	73%	0.41
	2	50	30	6	18	65%	0.31
	3	53	18	8	13	71%	0.32
<b>50 (77)</b>	1	29	34	14	0	37%	-0.41
	2	15	20	32	10	32%	-0.34
	3	32	2	20	23	71%	0.50
<b>40 (62)</b>	1	26	1	13	22	65%	0.28
	2	36	21	3	2	61%	0.01
	3	18	10	42	38	51%	0.1
<b>30 (46)</b>	1	18	9	6	12	66%	0.32
	2	18	8	12	8	56%	0.09
	3	14	4	15	13	58%	0.11

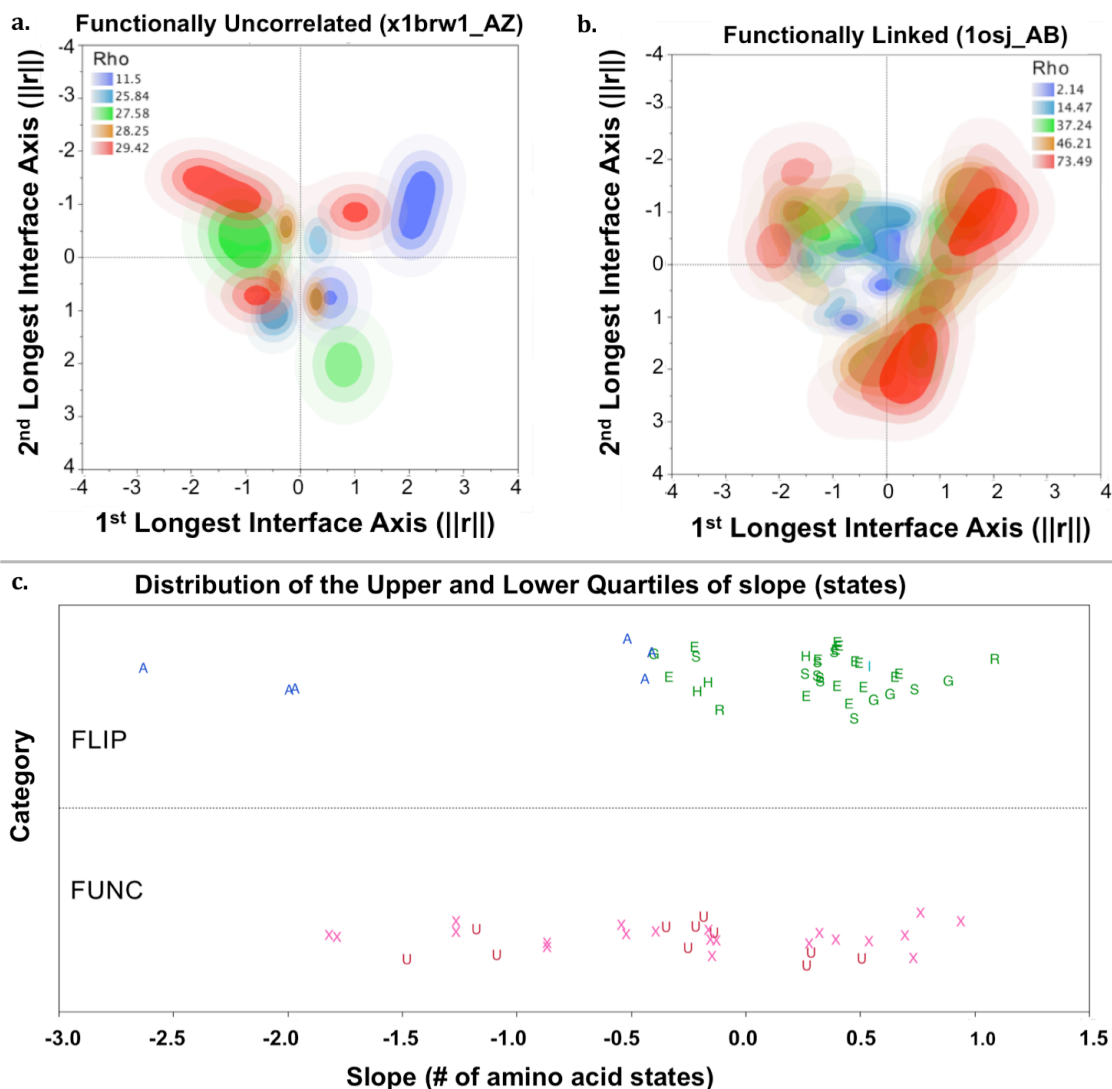
† TP: FLIP correctly predicted

† TN: FUNC correctly predicted

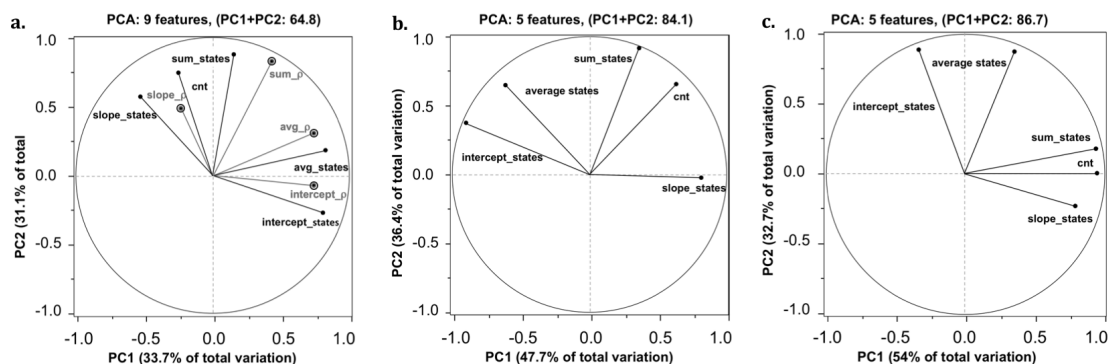
† FP: FUNC identified as FLIP

† FN: FLIP identified as FUNC

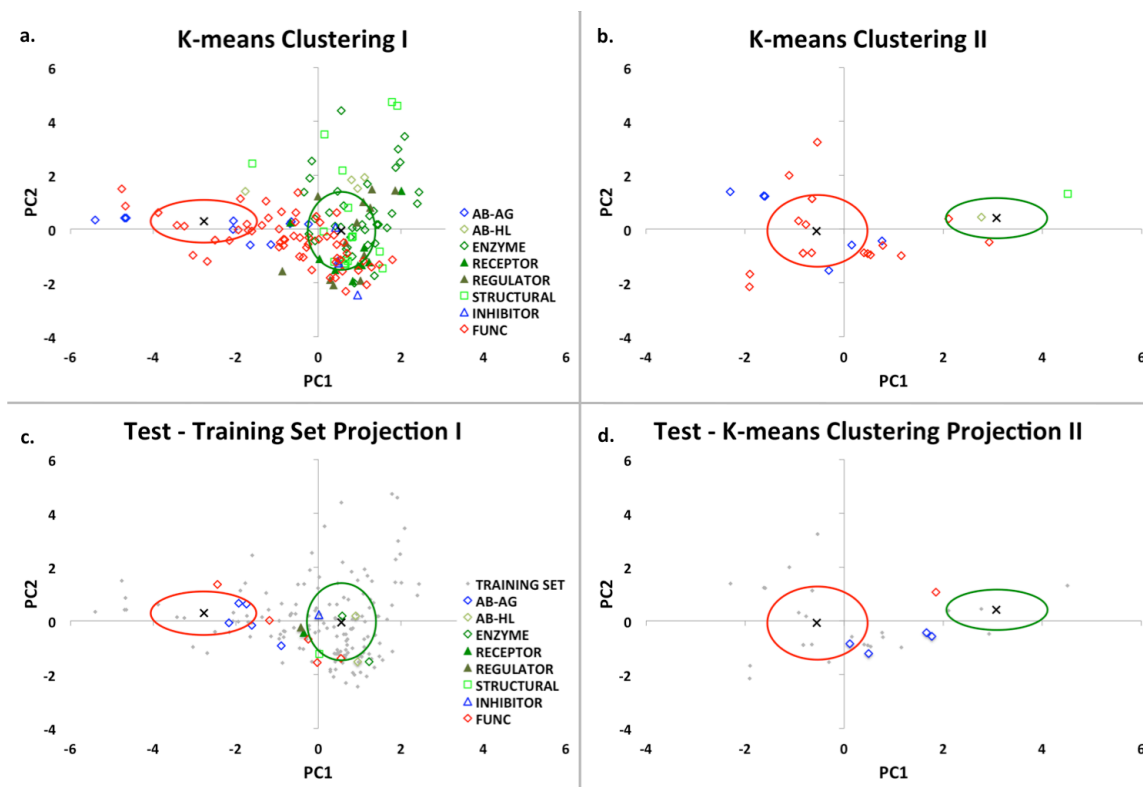
The distribution of overall accuracies and MCCs of repeated retraining when sub-samples of the training set were generated randomly in triplicate for subsets ranging from 90% to 20% of the original. The general accuracy is 67% until more than 50% of the training set is removed. MCCs stably range between approximately 0.07-0.32.



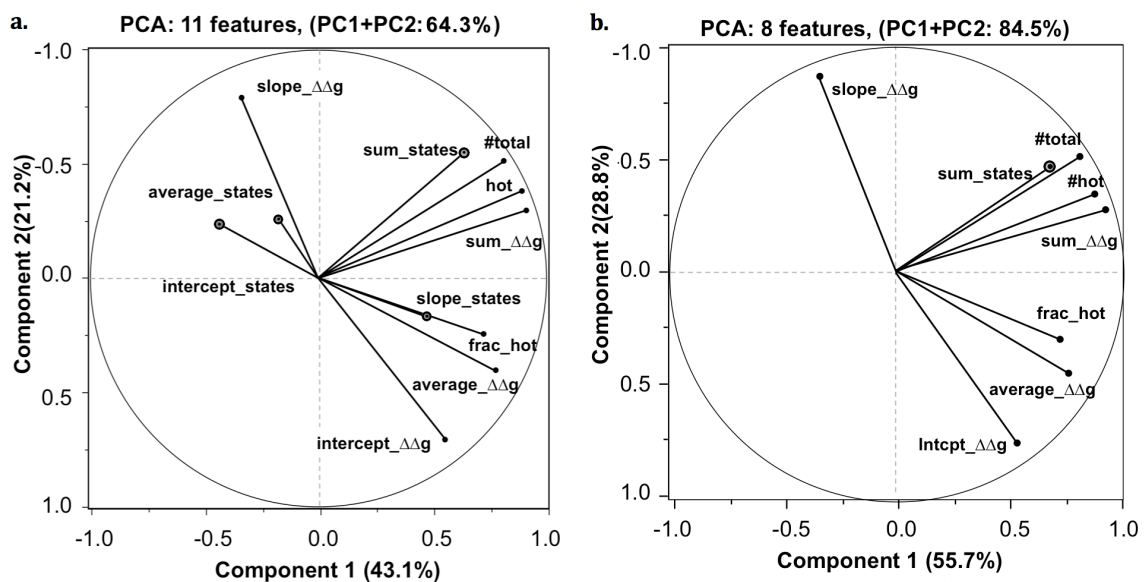
**Figure 3.1.** Distribution of conservation at interfaces. Figure 3.1a and b show histogrammed contour plots of the amino acid variation “Rho (r)” (blue: low variation, red: high variation). The plot axes are the first two principal components of the geometric distribution of  $C\alpha$  positions. PCA was used to align the interface along the X- and Y-axes. Amino acid variation in the crystal contact interface from PDBid: 1brw, chain A (x1brw1\_AZ) had more randomly distributed sites of conservation (a) while amino acid variation in the FLIP interface from PDBid: 1osj, chains A&B (b) shows a radial pattern of organization and a more conserved center. (c) Distribution of the top and bottom quartiles of the slopes from linear regressions of the amino acid variation with respect to the distance of each residue’s  $C\alpha$  from the mean of the  $C\alpha$  positions (Center of Interface). Green/Blue: FLIP. Red: FUNC. Symbols represent the functional sub-category of each interface: A=Ab-Ag, H=AbHL, E=Enzyme, R=Receptor, G=Regulator, I=Inhibitor, S=Structural, U=FUNC, X=XFunc.



**Figure 3.2.** Correlation of features with principal components. The loading plots (eigenvector co-efficients of each feature analyzed by PCA). The co-efficients indicate the contribution of each principal component. (a) All 9 features accounted for 64.8% of the total variation of the first two principal components. Features based on  $\rho$ , the Evolutionary Trace score (circles) had demonstrably smaller coefficients than features based on *states* (a count of the number of amino acid states at the position). (b) Removal of features based on  $\rho$ , produced a loading plot with PC 1 and 2 accounting for 84.1% of the resulting total variance. (c) Removal of 128 interfaces predicted to be FLIP in the first PCA, followed by a second round of PCA. This PCA reproduced 86.7% of the variance in the first two PCs.



**Figure 3.3.** PCA and K-means clustering of training and test sets. PCA and K-means were performed on the 154 interfaces in the FLIPdb. The 5 features identified in Figure 3.2 were used and the number of clusters was set to  $k = 2$ . Green (“cluster 1”) and red (“cluster 2”) represent 1 standard deviation for Euclidean distances around the cluster centroid marked by “x”. Interfaces are indicated with symbols representing their functional sub-category. Green and blue symbols are FLIP structures but blue symbols are specifically Ab-Ag and Inhibitor sub-categories. Red symbols are FunCs. (A) and (B): Training set. (C) and (D): Testing set. (A) 87 FLIP (92% of the FLIPS excluding most of the Ab-Ag) and 41 FunC interfaces were identified in cluster 1 (68% *precision*) (b) after removal of these 128 interfaces, a second analysis of the remaining 26 interfaces produced new clusters with 77% accuracy, 100% *precision*, and 73% *negative predication value*, but contributing little to the net accuracy. The overall accuracy across both (A) and (B) is 69%. (C) and (D) show the projection of the training set clusters on 18 unrelated PPIs.. The overall accuracy in both clusterings is 61%.



**Figure 3.4.** Correlation of evolution and energy features with principal components. The loading plots (eigenvector co-efficients of each feature analyzed by PCA) indicate the correlation of each variable to the principal components. (A) All 11 non-redundant energy and sequence conservation features accounted for 64.3% of the total variation. Overall, sequence-related terms (vectors with large circles) had much smaller coefficients than energy terms. (B) A new PCA with all features from (A) with coefficients less than 0.6 in both PC1 and PC2 removed generated a loading plot with new PC1 and PC2 accounting for 84.5% of the resulting 8-feature variance.

**Table S3.1.** FLIPdb-ET interface composition. Structures and interfaces used in the training and testing sets

PDBId	Chains	Category	Subcategory	# of Protein Chains	# of Interfaces	Citation
Training Set						
1acy	HP	FLIP	AbAg	3	3	1
1acy	LP	FLIP	AbAg	3	3	1
1adq	AH	FLIP	AbAg	3	3	2
1adq	AL	FLIP	AbAg	3	3	2
2bse	AE	FLIP	AbAg	5	9	3
2bse	AF	FLIP	AbAg	5	9	3
2bse	BD	FLIP	AbAg	5	9	3
2bse	BE	FLIP	AbAg	5	9	3
2bse	CD	FLIP	AbAg	5	9	3
2bse	CF	FLIP	AbAg	5	9	3
1acy	LH	FLIP	AbHL	3	3	1
1adq	LH	FLIP	AbHL	3	3	2
1gaf	LH	FLIP	AbHL	2	1	4
1tzi	AB	FLIP	AbHL	3	3	5
1wtl	AB	FLIP	AbHL	2	1	6
1b5e	AB	FLIP	ENZYME	2	1	7
1biq	AB	FLIP	ENZYME	2	1	8
1bis	AB	FLIP	ENZYME	2	1	9
1bjw	AB	FLIP	ENZYME	2	1	10
1bkp	AB	FLIP	ENZYME	2	1	11
1bmd	AB	FLIP	ENZYME	2	1	12
1brw	AB	FLIP	ENZYME	2	1	13
1bsl	AB	FLIP	ENZYME	2	1	14
1bsr	AB	FLIP	ENZYME	2	1	15
1cg2	AD	FLIP	ENZYME	2	2	16
1cg2	BC	FLIP	ENZYME	2	2	16
1cnz	AB	FLIP	ENZYME	2	1	17
1coz	AB	FLIP	ENZYME	2	1	18
1daz	CD	FLIP	ENZYME	4	3	19
1hjr	AC	FLIP	ENZYME	4	3	20
1hjr	BD	FLIP	ENZYME	4	3	20
1itv	AB	FLIP	ENZYME	2	1	21

livy	AB	FLIP	ENZYME	2	1	22
lnhk	RL	FLIP	ENZYME	2	1	23
lnsy	AB	FLIP	ENZYME	2	1	24
loro	AB	FLIP	ENZYME	2	1	25
losj	AB	FLIP	ENZYME	2	1	26
lpgt	AB	FLIP	ENZYME	2	1	27
lpky	AB	FLIP	ENZYME	4	4	28
lpky	AC	FLIP	ENZYME	4	4	28
lpky	BD	FLIP	ENZYME	4	4	28
lpky	CD	FLIP	ENZYME	4	4	28
lpre	AB	FLIP	ENZYME	2	1	29
lqks	AB	FLIP	ENZYME	2	1	30
lqr2	AB	FLIP	ENZYME	2	1	31
lr2f	AB	FLIP	ENZYME	2	1	32
lscu	AB	FLIP	ENZYME	4	5	33
lscu	DE	FLIP	ENZYME	4	5	33
lsft	AB	FLIP	ENZYME	2	1	34
lslt	AB	FLIP	ENZYME	2	1	35
lsmn	AB	FLIP	ENZYME	2	1	36
lsox	AB	FLIP	ENZYME	2	1	37
lubs	AB	FLIP	ENZYME	2	1	38
lvfr	AB	FLIP	ENZYME	2	1	39
xlubs1	BZ	FLIP	ENZYME	4	3	38
lcmi	AC	FLIP	INHIBITOR	4	3	40
lcmi	BD	FLIP	INHIBITOR	4	3	40
ldaz	CB	FLIP	INHIBITOR	4	3	19
ldaz	DA	FLIP	INHIBITOR	4	3	19
lppf	EI	FLIP	INHIBITOR	2	1	41
lcde	BA	FLIP	RECEPTOR	2	1	42
leaj	AB	FLIP	RECEPTOR	2	1	43
lfcc	AC	FLIP	RECEPTOR	2	1	44
lh0t	AB	FLIP	RECEPTOR	2	1	45
lnrv	AB	FLIP	RECEPTOR	2	1	46
2bse	AB	FLIP	RECEPTOR	5	9	3
2bse	AC	FLIP	RECEPTOR	5	9	3
2bse	BC	FLIP	RECEPTOR	5	9	3
2cii	AB	FLIP	RECEPTOR	4	2	47
2cii	AC	FLIP	RECEPTOR	4	2	47

1awi	AB	FLIP	REGULATOR	3	3	48
1awi	AP	FLIP	REGULATOR	3	3	48
1awi	BP	FLIP	REGULATOR	3	3	48
1edh	AB	FLIP	REGULATOR	2	1	49
1f13	AB	FLIP	REGULATOR	2	1	50
1f4v	AD	FLIP	REGULATOR	3	2	51
1psr	AB	FLIP	REGULATOR	2	1	52
1qfh	AB	FLIP	REGULATOR	2	1	53
1tx4	AB	FLIP	REGULATOR	2	1	54
2arc	AB	FLIP	REGULATOR	2	1	55
2trc	BG	FLIP	REGULATOR	3	2	56
2trc	BP	FLIP	REGULATOR	3	2	56
1cmb	AB	FLIP	STRUCTURAL	2	1	57
1f95	AB	FLIP	STRUCTURAL	4	1	58
1f95	AC	FLIP	STRUCTURAL	4	1	58
1f95	BD	FLIP	STRUCTURAL	4	1	58
1sa0	AB	FLIP	STRUCTURAL	5	6	59
1sa0	AE	FLIP	STRUCTURAL	5	6	59
1sa0	BE	FLIP	STRUCTURAL	5	6	59
1sa0	CD	FLIP	STRUCTURAL	5	6	59
1sa0	CE	FLIP	STRUCTURAL	5	6	59
1sa0	DE	FLIP	STRUCTURAL	5	6	59
1tub	AB	FLIP	STRUCTURAL	2	1	60
2bkh	AB	FLIP	STRUCTURAL	2	1	61
2bki	AB	FLIP	STRUCTURAL	2	1	61
3kin	AB	FLIP	STRUCTURAL	4	3	62
3kin	BD	FLIP	STRUCTURAL	4	3	62
3kin	CD	FLIP	STRUCTURAL	4	3	62
1atn	AD	FUNC	FUNC	2	1	63
1azt	AB	FUNC	FUNC	2	1	64
1bin	AB	FUNC	FUNC	2	1	65
1bkz	AB	FUNC	FUNC	2	1	66
1bo5	OZ	FUNC	FUNC	2	1	67
1c02	AB	FUNC	FUNC	2	1	68
1cmi	AB	FUNC	FUNC	4	3	69
1cqx	AB	FUNC	FUNC	2	1	70
1f4v	AB	FUNC	FUNC	3	2	51
1gqp	AB	FUNC	FUNC	2	1	71

1hjr	AB	FUNC	FUNC	4	3	20
1lcu	AB	FUNC	FUNC	2	1	72
1m6k	AB	FUNC	FUNC	2	1	73
1naw	AB	FUNC	FUNC	2	1	74
1nmt	AB	FUNC	FUNC	3	2	75
1nmt	BC	FUNC	FUNC	3	2	75
1oio	AB	FUNC	FUNC	2	1	76
1ome	AB	FUNC	FUNC	2	1	77
1scu	AE	FUNC	FUNC	4	5	33
1scu	BD	FUNC	FUNC	4	5	33
1scu	BE	FUNC	FUNC	4	5	33
1vio	AB	FUNC	FUNC	2	1	78
1xca	AB	FUNC	FUNC	2	1	79
3k3e	AB	FUNC	FUNC	2	1	80
x1b8e4	AE	FUNC	XFunc	2	1	81
x1brw3	BY	FUNC	XFunc	2	1	13
x1bsf8	AJ	FUNC	XFunc	2	1	82
x1bsr1	AB	FUNC	XFunc	2	1	15
x1dto6	AG	FUNC	XFunc	2	1	83
x1dv82	AC	FUNC	XFunc	2	1	84
x1e872	AC	FUNC	XFunc	2	1	85
x1erv2	AC	FUNC	XFunc	2	1	86
x1gaf1	HZ	FUNC	XFunc	2	1	4
x1gaf2	LY	FUNC	XFunc	2	1	4
x1gaf3	HL	FUNC	XFunc	2	1	4
x1ihk3	AD	FUNC	XFunc	2	1	87
x1neu4	AE	FUNC	XFunc	2	1	88
x1oal6	AG	FUNC	XFunc	2	1	89
x1oro1	AZ	FUNC	XFunc	2	1	25
x1oro3	AZ	FUNC	XFunc	2	1	25
x1ppf1	EI	FUNC	XFunc	2	1	41
x1ppf2	EZ	FUNC	XFunc	2	1	41
x1ppf3	EI	FUNC	XFunc	2	1	41
x1py94	AE	FUNC	XFunc	2	1	90
x1sox2	BY	FUNC	XFunc	2	1	37
x1tx41	BA	FUNC	XFunc	2	1	54
x1tx42	AZ	FUNC	XFunc	2	1	54
x1tx43	AB	FUNC	XFunc	2	1	54

x1uq54	AE	FUNC	XFunC	2	1	91
x1vfr2	AZ	FUNC	XFunC	2	1	39
x1vfr3	BY	FUNC	XFunC	2	1	39
x1wtl1	BZ	FUNC	XFunC	2	1	6
x1wtl2	BZ	FUNC	XFunC	2	1	6
x1xca1	AZ	FUNC	XFunC	2	1	79
x1xca2	BY	FUNC	XFunC	2	1	79
x1xca3	BA	FUNC	XFunC	2	1	79
x3bri2	AC	FUNC	XFunC	2	1	92
x3il85	AF	FUNC	XFunC	2	1	93
<b>*NOTE: Canonical XFunC PDBids are the 2nd-5th characters of their listed PDBids</b>						
Test set						
1avx	A:B	FLIP	Enzyme	2	1	94
1mah	A:F	FLIP	Inhibitor	2	1	95
1hiv	A:B	FLIP	Enzyme	2	1	96
1kac	A:B	FLIP	Receptor	2	1	97
1buh	A:B	FLIP	Regulator	2	1	98
1ijj	A:B	FLIP	Structural	2	1	99
1ahw	A:B	FLIP	AbHL	2	1	100
1ahw	A:C	FLIP	AbAg	2	1	100
1ahw	A:F	FLIP	AbAg	2	1	100
1ahw	B:C	FLIP	AbAg	2	1	100
1ahw	D:E	FLIP	AbHL	2	1	100
1ahw	D:F	FLIP	AbAg	2	1	100
1ahw	E:F	FLIP	AbAg	2	1	100
x1kac1	B:A	FunC	XFunC	2	1	97
x1avx1	A:A	FunC	XFunC	2	1	94
x1mah1	F:A	FunC	XFunC	2	1	95
x1buh1	B:A	FunC	XFunC	2	1	98
x1ijj1	B:A	FunC	XFunC	2	1	99
<b>*NOTE: Canonical XFunC PDBids are the 2nd-5th characters of their listed PDBids</b>						

The FLIPdb-ET database contained 154 pairwise PPIs. The Test set contained 18 pairwise PPIs.

## CHAPTER IV

### ENERGY CENTRALITY RELATIONSHIP REDUCES FALSE POSITIVE PREDICTION IN PROTEIN DOCKING

#### **Abstract**

Interacting protein networks are responsible for a multitude of biological functions and molecular docking has become an important tool in predicting protein interactions. A common problem with docking is the generation of a large number of false positives. In previous work, we have used our Energy Centrality Relationship (ECR) concept to differentiate between specific interactions between two categories of interfaces, Functionally-Linked Interacting Proteins (FLIPs) and Functionally uncorrelated Contacts (FunCs). We found that the positional and energetic correlation patterns arising from ECR can discriminate FLIP/FunC. Here we test ECR's ability to identify near-native ( $\leq 5$  Å RMSD) poses arising from docking components of known protein complexes. After generating docking decoys for structures of representatives of diverse protein functional categories using the docking software HEX, we demonstrate ECR reduces false positives in quaternary structure prediction. The ECR methodology was able to predict near-native poses in 50% of the cases, representing an increase of 9% relative to HEX alone.

## Introduction

Proteins participate in various cellular processes that are required for biological systems. Various groups have studied proteins interactions<sup>1-3</sup>; however, predicting interfaces of complex proteins via experimental analysis is often difficult and time-consuming<sup>4</sup>. Understanding the contribution of an interface to the stability of a complex requires knowledge of the three-dimensional structures of proteins. Classical methods, such as X-ray crystallography and NMR (nuclear magnetic resonance) spectroscopy, produce high-resolution three-dimensional structures. However, due to the laborious nature of experimental protein structure generation, computational methods, like protein-protein docking, can be used as a structure-based approach for protein interaction prediction utilizing knowledge from existing three-dimensional protein structures.

Docking algorithms employ a wide variety of strategies to identify near-native quaternary conformations (poses) of receptor and ligand (interactors). Some of these search methods include: an exhaustive global search, particularly Fast Fourier Transform (FFT) based methods<sup>5-14</sup>, local shape matching approaches<sup>15</sup> and randomized search<sup>16-20</sup>. A majority of docking algorithms use FFT-based<sup>21</sup> approaches as they are both well-optimized for quaternary structure prediction and parallelized for significant speed up of related calculations<sup>6, 9, 21</sup>. These programs are fast and effective in interaction prediction; however, most of these methods consider the interactors as rigid bodies<sup>9, 11, 12, 14, 22</sup>. Proteins, however, are dynamic and the enhanced computational complexity of adding flexibility to increase accuracy generally decreases computational

through-put in ways that are challenging to compensate for using FFT-based methods. As such, Rigid-body docking is often still used both as an end-goal and also to limit the conformational search space analyzed in flexible docking methods<sup>23</sup>.

Though docking is used to predict Protein-Protein Interaction (PPI) structures, the validity and efficacy of this method can be evaluated by re-docking known quaternary interfaces. Since the starting structure for such re-docking is obtained from an experimentally determined structure, for docking purposes that experimental structure is considered to be the “native” conformation. However, protein structure files obtained from the Protein Databank (PDB)<sup>24</sup> contain the atomic coordinates of the asymmetric unit or ASU. The ASU is the smallest portion of a crystal that can be used to construct a repeating unit of the crystal and it may or may not be a biologically relevant conformation. Contacts between neighboring ASUs are often observed in X-ray structures<sup>25</sup>. These ‘crystal contacts’ can represent the biologically functional state of the protein, though generally they do not<sup>25</sup>. Therefore, within a crystal structure, two monomers could form both functional and non-functional interfaces. It can be challenging to distinguish between the two<sup>26-30</sup>. In previous research, we have defined functionally relevant interfaces as FLIP (Functionally-Linked Interfaces of Proteins) and interfaces that do not contribute to function as FunC (Functionally uncorrelated Contacts)<sup>31</sup>. In our work, we use this operational definition as the work of others on PPI detection often characterizes the occurrence or existence of a PPI without considering whether that PPI is critical to biological function<sup>32, 33</sup>.

A docking algorithm's accuracy depends on its ability to distinguish between native-like and non-native conformations. Here a native-like conformation is defined as any structure within 5 Å root mean square deviation (RMSD) of the known structure. The accuracies of various docking software can be greatly improved by using knowledge-based scoring functions<sup>6, 34-38</sup> to generate more native-like conformations. The absence of *a priori* information of potential binding sites requires extensive sampling to identify possible low-energy conformations. Protein properties that tend to correlate with native interfaces and thus are useful to interface prediction are shape complementarity<sup>5, 11, 14, 22, 39, 40</sup>, residue hydrophobicity<sup>41, 42</sup>, residue conservation at the interface<sup>43</sup>, and buried surface area upon complexation<sup>44</sup>. Scoring terms based on system energies, including van der Waals, electrostatics, and solvation energies help in evaluating poses by calculating the number of favorable intermolecular interactions and have been seen to improve the success rate of docking algorithms<sup>14, 22, 45</sup>. Even though some properties have significant correlations with known interfaces, it is still difficult to fully distinguish between native and non-native interfaces using a single feature. Many algorithms therefore combine analysis of several different features to improve docking prediction<sup>46-48</sup>.

Despite various developments in docking, the common problem of generating a high number of false positive poses remains. One way to approach this problem is to develop a post-filter for docking that can identify a few top structures using features that are known to be important to interface maintenance<sup>49, 50</sup>. In our earlier works, we investigated the use of both sequence conservation and energy related terms to

differentiate between FLIPs and FunCs<sup>31, 51</sup>. While interface sequence conservation can distinguish FLIP from FunC with an accuracy of 69%, the use of energy related terms could also distinguish the groups but with a better prediction accuracy of 77%.

We hypothesize that, if the interfaces generated by protein docking (“poses”) are false positives or aggregates, they would exhibit physical and biochemical properties more similar to known FunCs, while those with near native specific interactions would exhibit properties similar to known FLIPs. We suggest that post-docking analyses of these properties using our ECR methodology will improve the overall prediction accuracy.

## Results

### *ECR prediction of docking targets*

The results of docking followed by ECR analysis of the 22 proteins in FLIPdb-lite are shown in Table 4.1. FLIPdb-lite is a derivative of the FLIPdb from our previous work (see Methods). These 22 interfaces were identified as *easy*, *intermediate*, and *hard* targets, where *easy* interfaces were defined as interfaces that upon docking generated greater than 75% native-like poses, *hard* generated fewer than 5% native-like poses, and *intermediate* generated between 5%-75% native-like poses. The Principal Component (PC) projections (see Methods) of all poses were identified as presenting in either the FLIP or FunC prediction clusters. For each interface, the RMSD of HEX’s top scoring pose (“HEX”) as well as the pose with the lowest RMSD from the original structure (“HEX best”) were compared with ECR’s top prediction (defined as the pose with the most positive PC1, in the PCA). Also in Table 4.1 is the structure with the smallest

RMSD within the FLIP cluster (“ECR best RMSD”). The FLIP or FunC cluster analyses of three representative examples (one from each class of target: PDBids *Icdc*, *Ibsr* and *Itub*) are shown in Figure 4.1.

A summary comparison of HEX and ECR’s overall performance is shown in Table 4.2. HEX’s success at predicting native-like structures when considering only its top solution (best scoring pose) was 41% (9 of 22). Post-filtering with ECR, however, increased success by 2 to 11 of 22 interfaces (50%), a 9% increase in accuracy. If all poses generated by HEX regardless of their score were considered, HEX achieved a success of one additional interface (12 of 22, 54%). Since our current ECR selection method identifies just one structure, its success remained at 50% (11 of 22). Dockings of AbAg (Antibody-Antigen) and Enzyme-Inhibitor interactions were particularly challenging (low success rate in both HEX and ECR). As such, ECR post-filtering offered no additional benefits for such interfaces. Regardless, post-filtering with ECR improved docking by 9% suggesting ECR is accounting for principles not represented by HEX’s docking method alone.

In addition, Table 4.3 shows our success of prediction based on the analysis of distribution plots of RMSD values for all generated poses (Figures 4.2 and 4.S2). Of the 18 FLIPs analyzed, 13 (70%) identified a pose in the lowest quartile of the RMSD distribution. Of these, 10 were correctly predicted (near-native). Our accuracy of prediction from the lower quartile was thus 77% (10/13). This could be of particular importance in our effort to post-filter docking poses as this suggests a strategy of

eliminating all poses not in the lowest quartile of produced poses. Such a strategy would therefore be predicted to eliminate the majority of the false positive poses from a docking run.

### ***Representative conformations***

#### *Easy conformation: Icdc*

*Icdc* is a dimeric membrane glycoprotein that functions as a cell adhesion molecule. All poses generated were native-like. Clustering analysis of *Icdc* shows that almost all structures generated in the docking run (98%) partitioned in the FLIP cluster region (Figure 4.1a). The structure identified by ECR was solution 41 from HEX, which had an RMSD of 0.86Å from the experimental structure. The “best” solution and HEX’s pick was solution 1 with an RMSD of 0.19 Å.

#### *Intermediate conformation: Ibsr*

*Ibsr* is the structure of the homodimeric enzyme, bovine seminal ribonuclease. 25% of poses generated were native-like but very few partition in the FLIP region (Figure 4.1b). The structure identified after post-filtering by ECR was solution 3 from HEX, which had an RMSD of 0.21Å. This was also HEX’s lowest RMSD structure. The reduction of the number of possible poses from 25% (251/1000) to 1% (11/1000) demonstrates the ability of ECR as a beneficial post-filter.

### *Hard conformation: Itub*

*Itub* is a heterodimer of  $\alpha\beta$  tubulin, the microtubule subunit. No poses generated by HEX were initially thought to be native-like. Since the initial PCA-projection analysis did not identify any poses in the FLIP region, features were re-analyzed using the second PCA-projection criteria from our prior ECR work (Figure 4.1c and d). The structure subsequently identified by ECR was solution 153 from HEX, which had an RMSD of 78.14 Å. On further analysis of solution 153, we found that it had an RMSD of 2.2Å when superposed on the crystal symmetry transformed conformation of *Itub* (RMSD of 2.2Å) shown in Figure 4.3b. As this crystal symmetry conformation is that of a separate but related microtubule conformation ( $\beta\alpha$  tubulin), we take this to demonstrate the ability of the ECR method to identify native poses.

### ***Distribution of poses***

Histogram and box plots of RMSDs of poses relative to the known structure are shown in Figure 4.2. HEX's ability to generate native-like poses varies substantially for the three representative conformations discussed above. While HEX produced a slightly more native-like structure for *Icdc*, ECR's result was still in the lower quartile (which includes the smallest RMSD structures) and was nearly a low RMSD outlier. HEX did not produce nearly as many native-like structures for *Ibsr* as it did for *Icdc* (25% and 100%, respectively); however ECR's prediction was still in the lower quartile. *Ibsr*'s poses also demonstrated two distinct distributions of RMSD, indicating HEX identified

two groups of related, energetically favorable, quaternary conformations, only one of which was native-like. Of these two, ECR was successful at identifying the near native conformation. All of HEX's poses for *Itub* had large RMSDs from the source native structure, the smallest being 30.25 Å. Despite there being no obvious near-native poses in this distribution, ECR identified the symmetry-related native structure of *Itub*. Histograms and box plots of additional interfaces are shown in supplementary data, Figure 4.S2.

### ***Symmetry operator prediction***

Another successful example is *xIubsI*, a crystal symmetry operator defined structure generated using the symmetry operations of the PDB structure *Iubs* (chains A and B). The native structure of *Iubs* includes a two-chain hetero-dimer (Figure 4.4a). *xIubsI* consists of two B chains (Figure 4.4b) and as such this interface was not expected to be functional. However, after FLIPdb creation and analysis, additional literature information was identified suggesting that active *Iubs* forms a dimer of dimers that creates a functional B:B interface (Figure 4.4d). Interestingly, native-like poses of this structure were identified both by ECR (solution 2, Figure 4.4c) and HEX (solution 1, not shown as it is not easily visually distinguishable from ECR's pick). The RMSD between HEX solution 1 and *xIubsI* was 1.1 Å, while ECR's pick had an RMSD of 0.33 Å (Figure 4.4e).

## Discussion

### *Docking poses exhibit different distribution patterns*

The ECR training set demonstrates specific distribution patterns corresponding with different categories of function<sup>31</sup>. Analysis of the ECR training set showed that FLIPs were mostly clustered in the positive PC1 region while FunCs mostly occupied the negative PC, quadrants 1 and 2 respectively. The projection of docking poses on the ECR training set showed that the docked poses were mostly distributed in the negative PC1 region, the region generally occupied by FunC interfaces. One exception was *Icdc* (a receptor, Figure 4.1a), where most poses were distributed in the positive PC1 region. These differences in the general pattern of FLIP, FunC, and docking pose distribution suggest a fundamental difference in the physical chemistry of these different types of interface structures. The poses generated by HEX are the top scoring predictions from the docker and as such are extremely unlikely to be random predictions. Still, the generation of a high number of false positive poses not placed in either FunC or FLIP cluster indicates that HEX possibly predicts complexes with high affinity, which is important to the stability of the complex. However, high affinity alone does not necessarily mean high specificity<sup>29, 31</sup>.

### *ECR can recognize symmetry-related native poses*

ECR analysis of the *Itub* and *xIubs1* structures identified native poses that would otherwise have been missed or rejected. *Itub* is a hetero-dimer that is a component of the

microtubule filaments within the cytoskeleton. The structure of a microtubule filament involves spirally repeating units of this hetero-dimer. As such, each monomer could have multiple FLIP interfaces, not simply the one observed in the crystal structure. Initially, the docking analysis of *ltub* did not appear to generate any native-like poses nor was a pose identified in the FLIP cluster in the first clustering projection. The second clustering projection identified 11 possible interfaces in the FLIP region. Inspection of the pose with the most positive PC1 value revealed that this pose was native-like (2.2 Å RMSD) relative to a symmetry-related FLIP interface in the filament.

The other example, *xlubs1*, was initially identified as an XFunC as this structure was generated using the crystal symmetry operations of *lubs*, an enzyme. Since *xlubs1* was initially misclassified, this structure serves as an, albeit unintentional, negative control of the ECR methodology. Docking was not expected to generate a native-like A:B pose like that in *lubs*, since *xlubs1* was a symmetry operator related pose involving only the B chain of *lubs* in a B:B interface. However, ECR identified 6 docked poses within 5 Å of *xlubs1*. *lubs* was found in actuality to be a two-chain hetero-tetramer and, in creating *xlubs1*, we had accidentally created an alternate FLIP. As such, ECR's best solution was also an alternate FLIP.

These two examples indicate that ECR is capable of identifying the underlying principles of functional linkage in protein association. It is possible that ECR identified a combination of features related to sensitivity to mutation that was not accounted for in

HEX. Some of this sensitivity comes from local structure perturbation due to changes in side-chain packing.

HEX could possibly get the correct answer by generating more poses and selecting a larger number of high scoring structures but that would be more computationally expensive since it is difficult to predict the minimum number of poses required to find the best solution. The goal in this work is to filter existing docking poses to predict the most native-like structure. As such, generating more poses, while possibly providing additional accuracy, would also dramatically increase computational demands.

HEX's larger success came at the cost of generating hundreds of alternative poses (13% increase in prediction success when 1000 poses were considered as opposed to just the best scoring pose). Most of this expensive but improved success was due to correctly predicting an enzyme-inhibitor interface PDB: 1ppf (ECR has very limited success in inhibitor prediction) and a FunC, PDB: 1bin (the current ECR selection criteria focuses on FLIP, not FunC, identification). AbAg and Enzyme-Inhibitor interfaces were especially difficult to predict by HEX as these surfaces have low shape complementarity, a key feature in interface prediction by docking. ECR analysis of these types of interfaces revealed that they, in general, partition with FunCs in the PCA suggesting they may be governed by similar underlying principles as FunCs or at the very least evolve via alternative mechanisms than other FLIPs.

### ***HEX struggles with AbAg***

As stated above, the binding region of AbAgs tends to be flatter than other FLIP interfaces such as enzymes<sup>42</sup>. This difference in interface structure is a problem for most docking algorithms that use complementarity as a key factor in interface detection. This can partly be overcome by using docking algorithms such as Zdock<sup>14</sup> or PatchDock<sup>15</sup> that either mask the non-interacting regions of the proteins and allow conformational sampling only in the binding area or use a dedicated docking algorithm<sup>52</sup>. The poor performance of AbAg could also be because of enhanced sequence variability at the interface that obscures sequence conservation signals<sup>53</sup>.

### **Conclusion**

Docking analysis coupled with ECR showed an improvement over docking alone, suggesting ECR accounts for principles not represented by docking software. This post-filtering approach was less successful for AbAg and Enzyme-Inhibitor interfaces, perhaps due to different or limited evolutionary pressures, resulting in more FunC-like features. Overall, use of ECR as a docking post-filter not only reduces false-positive prediction by 9% but also recognizes false negatives in our bound docking studies. In addition, it helps identify when no true-positives are likely to have been produced. Finally, ECR identifies patterns of physico-chemical properties that help distinguish affinity from specificity in protein association.

## Methods

### *Dataset: FLIPdb – lite*

For construction of FLIPdb, please refer to Sudarshan et.al<sup>31</sup>. Here we analyzed a truncated version of FLIPdb we called FLIPdb-lite. To construct FLIPdb-lite we identified 3 proteins from each protein sub-category of FLIPdb. The Euclidean distance from the values of 7 energy and geometry features of all FLIPdb proteins from FLIP and FunC cluster centroids have previously been determined<sup>31</sup>. The interfaces of each sub-category were selected by determining the 2 within that sub-category with the closest and the 1 with the farthest 7-feature Euclidean distance to the appropriate FLIP or FunC centroid. If this procedure selected an additional interface from a previously identified PDB structure, this new interface was rejected and the interface with the next closest or farthest feature distance was chosen. The selected proteins that generated docking poses were included in FLIPdb-lite.

### *Molecular docking*

The protein docking program HEX was used to generate poses for protein structures in FLIPdb-lite. Shape and electrostatic contributions to the docking correlation were enabled. The program was set to generate 1000 poses at Euler rotational increments of 7 degrees and a twist of 2 degrees. Chain 1 of the dimer was always identified as the receptor, which was fixed and chain 2 was identified as the ligand, which sampled

conformational space around the receptor. For each docking pose generated, an all atom RMSD to the PDB structure was calculated.

### ***Computational alanine scanning (CAS)***

The CAS method of Kortemme and Baker<sup>2, 54</sup> was used to process all the interfaces in the FLIPdb. In brief, this method evaluates enthalpy and free energy of solvation terms over conformations arising from a rotamer library for both the existing and alanine substituted residues in a PPI (native Gly and Pro excluded). These terms are used to determine a pseudo-free energy change upon substitution ( $\Delta\Delta G$ )<sup>54</sup>

### ***Interfacial geometry***

Interfacial residues were defined using the same interface definition as in the CAS method of Kortemme and Baker<sup>54</sup>. The geometric distribution of residues in each PPI were determined by calculating the displacement ( $\Delta r$ ) of the C $\alpha$  position from the mean of the C $\alpha$  positions (termed the Center of Interface, CoI) using software written by the authors. A linear regression of the  $\Delta\Delta G$  and  $\Delta r$  data to a first-order polynomial ( $\Delta\Delta G = \text{slope} * \Delta r + \text{intercept}$ ) was calculated for each interface using software written by the authors as well as GNUPLOT<sup>55</sup>

### ***Energetic and geometric features***

Calculations used in this work followed our previous protocol<sup>31</sup>, identifying 7 features for each interface: the slope (slope\_ $\Delta\Delta G$ ), intercept (intcpt\_ $\Delta\Delta G$ ), net sum of all

$\Delta\Delta G$  changes (sum\_ $\Delta\Delta G$ ), mean  $\Delta\Delta G$  for all interface residues (avg\_ $\Delta\Delta G$ ), total number of residues in the interface (#total), number of residues with  $\Delta\Delta G$  larger than +1 kcal/mol (#hot), and the ratio of “hot” to total (frac\_hot).

### ***Principle component analysis (PCA)***

In our prior work, Principal Component Analysis of the variation of CAS energetic and geometric feature data for a training set of PPI was undertaken<sup>31</sup>. PCA determines a set of linearly-uncoupled eigenvectors from normalized correlations between variables that progressively describe the largest sources of variance in a data set<sup>56</sup>

### ***K-means clustering***

K-means clustering<sup>57</sup> is a data analysis method that clusters observations into a specific number of clusters by attempting to find the point(s) that have the lowest mean variation from the other input data. When combined with PCA, the combination of features that allows input data to be clustered can be identified. In this work, we projected new energy and distance data for the poses of the 22 proteins through a polynomial obtained from the PCA and K-means clustering in our earlier work on a training set of proteins<sup>31</sup>. From this projection, we were able to identify where along the 1<sup>st</sup> two Principal Component vectors a new protein would present, as well as whether that protein would occur in a FLIP or FunC predicted cluster. We defined ECR’s pick from all

available poses to be the structure with the most positive PC1 in the FLIP region of our clustering analysis.

## References

1. Szilágyi A, Grimm V, Arakaki AK, Skolnick J (2005) Prediction of physical protein-protein interactions. *Phys Biol* 2:S1-S16.
2. Kortemme T, Baker D (2002) A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A* 99:14116-14121.
3. Jones S, Thornton JM (1995) Protein-protein interactions: a review of protein dimer structures. *Prog Biophys Mol Biol* 63:31-65.
4. Fink F, Ederer S, Gronwald W (2009) Protein-Protein Interaction Analysis by Docking. *Algorithms* 2:429.
5. Eisenstein M, Shariv I, Koren G, Friesem AA, Katchalski-Katzir E (1997) Modeling supra-molecular helices: extension of the molecular surface recognition algorithm and application to the protein coat of the tobacco mosaic virus. *J Mol Biol* 266:135-143.
6. Gabb HA, Jackson RM, Sternberg MJ (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* 272:106-120.
7. Jackson RM, Gabb HA, Sternberg MJ (1998) Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J Mol Biol* 276:265-285.
8. Vakser IA (1997) Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins Suppl* 1:226-230.
9. Ritchie DW, Kemp GJ (2000) Protein docking using spherical polar Fourier correlations. *Proteins* 39:178-194.
10. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Improved protein-ligand docking using GOLD. *Proteins* 52:609-623.
11. Kozakov D, Brenke R, Comeau S, Vajda S (2006) PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins JID* - 8700181.
12. Li L, Guo D, Huang Y, Liu S, Xiao Y (2011) ASPDock: protein-protein docking algorithm using atomic solvation parameters model. *BMC Bioinformatics* 12:36-36.

13. Palma PN, Krippahl L, Wampler J, Moura J (2000) BiGGER: a new (soft) docking algorithm for predicting protein interactions. *Proteins JID* - 8700181.
14. Chen R, Li L, Weng Z (2003) ZDOCK: an initial-stage protein-docking algorithm. *Proteins* 52:80-87.
15. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* 33:W363-W367.
16. Fernández-Recio J, Totrov M, Abagyan R (2003) ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins* 52:113-117.
17. de Vries S,J., van Dijk A,D.J., Krzeminski M, van Dijk M, Thureau A, Hsu V, Wassenaar T, Bonvin AMJJ (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins* 69:726-733.
18. Zacharias M (2005) ATTRACT: protein-protein docking in CAPRI using a reduced protein model. *Proteins* 60:252-256.
19. Dominguez C, Boelens R, Bonvin AMJJ (2003) HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information. *J Am Chem Soc* 125:1731-1737.
20. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol* 331:281-299.
21. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A* 89:2195-2199.
22. Ten Eyck LF, Mandell J, Roberts VA, Pique ME (1995) Surveying Molecular Interactions with DOT. *Supercomputing, 1995 Proceedings of the IEEE/ACM SC95 Conference*:22-22.
23. Moreira IS, Fernandes PA, Ramos MJ (2010) Protein-protein docking dealing with the unknown. *J Comput Chem* 31:317-342.
24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235-242.

25. Levy ED (2007) PiQSi: protein quaternary structure investigation. *Structure* 15:1364-1367.
26. Henrick PH (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins* 41:47-57.
27. Søndergaard C,R., Garrett AE, Carstensen T, Pollastri G, Nielsen JE (2009) Structural artifacts in protein-ligand X-ray structures: implications for the development of docking scoring functions. *J Med Chem* 52:5673-5684.
28. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2004) A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol* 336:943-955.
29. Henrick KE (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 372:774-797.
30. Harris R, Olson AJ, Goodsell DS (2008) Automated prediction of ligand-binding sites in proteins. *Proteins* 70:1506-1517.
31. Sudarshan S, Kodathala SB, Mahadik AC, Mehta I, Beck BW (2014) Protein-protein interface detection using the energy centrality relationship (ECR) characteristic of proteins. *PLoS One* 9:e97115-e97115.
32. Janin J, Rodier F (1995) Protein-protein interaction at crystal contacts. *Proteins JID -* 8700181.
33. Carugo O, Argos P (1997) Protein-protein crystal-packing contacts. *Protein Science : A Publication of the Protein Society JID -* 9211750.
34. Gohlke H, Hendlich M, Klebe G (2000) Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* 295:337-356.
35. Warren GL, Andrews CW, Capelli A, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, et al (2006) A critical assessment of docking programs and scoring functions. *J Med Chem* 49:5912-5931.
36. Muegge I, Martin YC (1999) A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem* 42:791-804.
37. Ishchenko AV, Shakhnovich EI (2002) SMoG2001 (SMoG2001): an improved knowledge-based scoring function for protein-ligand interactions. *J Med Chem* 45:2770-2780.

38. Krüger D,M., Ignacio Garzón J, Chacón P, Gohlke H (2014) DrugScorePPI knowledge-based potentials used as scoring and objective function in protein-protein docking. *PLoS One* 9:e89466-e89466.
39. Ritchie DW, Kemp GJ (2000) Protein docking using spherical polar Fourier correlations. *Proteins* 39:178-194.
40. Li L, Chen R, Weng Z (2003) RDOCK: Refinement of rigid-body protein docking predictions. *Proteins: Structure, Function, and Bioinformatics* 53:693-707.
41. Young L, Jernigan RL, Covell DG (1994) A role for surface hydrophobicity in protein-protein recognition. *Protein Sci* 3:717-729.
42. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. *J Mol Biol* 285:2177-2198.
43. Panchenko AR, Kondrashov F, Bryant S (2004) Prediction of functional sites by analysis of sequence and structure conservation. *Protein Science* 13:884-892.
44. Mishra S (2012) Computational prediction of protein-protein complexes. *BMC Res Notes* 5:495-495.
45. Garzon JI, López-Blanco JR, Pons C, Kovacs J, Abagyan R, Fernandez-Recio J, Chacon P (2009) FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics* 25:2544-2551.
46. Neuvirth H, Raz R, Schreiber G (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol* 338:181-199.
47. Bordner AJ, Abagyan R (2005) Statistical analysis and prediction of protein-protein interfaces. *Proteins* 60:353-366.
48. Hsin K, Ghosh S, Kitano H (2013) Combining machine learning systems and multiple docking simulation packages to improve docking prediction reliability for network pharmacology. *PLoS One* 8:e83922-e83922.
49. Talavera D, Robertson D, Lovell S (2011) Characterization of protein-protein interaction interfaces from a single species. *PloS One* JID - 101285081.
50. Brown NG, Chow D, Ruprecht K, Palzkill T (2013) Identification of the beta-lactamase inhibitor protein-II (BLIP-II) interface residues essential for binding affinity and specificity for class A beta-lactamases. .

51. Sudarshan S, Beck B (In Revision) <br />Functional Classification of Protein Interactions using Interface Spatial Distribution of Evolutionary Criteria. .
52. Brenke R, Hall DR, Chuang G, Comeau SR, Bohnuud T, Beglov D, Schueler-Furman O, Vajda S, Kozakov D (2012) Application of asymmetric statistical potentials to antibody-protein docking. *Bioinformatics* 28:2608-2614.
53. Lopes A, Sacquin-Mora S, Dimitrova V, Laine E, Ponty Y, Carbone A (2013) Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information. *PLoS Comput Biol* 9:e1003369-e1003369.
54. Kim KT (2004) Computational alanine scanning of protein-protein interfaces. *Sci STKE* 2004:pl2-pl2.
55. Williams T, Kelley C (2011) .
56. Pearson K (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* 2 11:559-572.
57. Hartigan JA (1973) Clustering. *Annu Rev Biophys Bioeng* 2:81-101.

**Table 4.1. Docking Results of FLIPlite**

PDB ID	Category	HEX RMSD (Soln #1)	HEX best RMSD (Soln #)	ECR RMSD (Soln #)	ECR best RMSD (Soln #)
1adq_AL	AbAg	86.83 (1)	32.03 (269)	72.86 (522)	32.03 (269)
1tzi_AV	AbAg	52.53 (1)	33.09 (137)	43.94(257)	‡
1adq_HL	Ab-HL	24.15 (1)	0.3 (2)	1.21 (11)	0.3(2)
1tzi_AB	Ab-HL	0.64 (1)	0.64 (1)	0.64 (1)	0.64(1)
1bsr_AB	Enzyme	0.63 (1)	0.21 (3)	0.21 (3)	0.21(3)
1bsl_AB	Enzyme	0.70 (1)	0.70 (1)	0.70 (1)	0.70(1)
1biq_AB	Enzyme	0.65 (1)	0.65 (1)	0.65 (1)	0.65(1)
3kin_BD	Structural	50.55 (1)	27.5 (596)	39.8 (241)	38.65(90)
1tub_AB	Structural	42.63 (1)	30.25 (179)	78.14/2.26* (153*)	30.25(179)
2bkh_AB	Structural	37.31 (1)	20.45 (996)	67.56 (395)	66.12(426)
1cdc_BA	Receptor	0.19 (1)	0.19 (1)	0.86 (41)	0.19 (1)
1cmb_AB	Regulator	0.32 (1)	0.32 (1)	0.32 (1)	0.32 (1)
1tx4_AB	Regulator	0.41 (1)	0.41 (1)	0.74(4)	0.41(1)
1awi_AB	Regulator	44.28 (1)	22.88 (770)	54.93(321)	‡
1ppf_EI	Inhibitor	17.03 (1)	2.73(5)	46.69(47)	10.44(560)
1cmi_BD	Inhibitor	0.46 (1)	0.33 (12)	0.84 (9)	0.33(12)
1daz_CB	Inhibitor	18.14 (1)	13.63 (984)	21.09(28)	‡
1bin_AB	FunC	9.51 (1)	3.18 (21)	23.09 (3)	12.38(61)
1cqx_AB	FunC	98.2 (1)	34.22 (543)	98.2(1)	‡
1c02_AB	FunC	38.65 (1)	10.94 (73)	39.11 (46)	10.94 (73)
x1ppf3_EI	XFunC	52.01 (1)	14.0 (753)	55.48 (60)	24.88(472)
x1ubs1_BZ	XFunC/FLIP	1.10 (1)	0.33 (2)	0.33 (2)	0.33 (2)

\*Superposition of HEX solution 153 on the crystal symmetry transformed conformation of *1tub* had an RMSD of 2.2Å

‡No poses were identified in the FLIP region.

**Table 4.1.** *Docking Results of FLIPlite*

Table with HEX and ECRs best prediction for each interface in FLIPlite. The RMSD of the highest scoring structure (HEX's #1 solution); HEX solution with the smallest RMSD (HEX best) and its solution number (HEX Soln #); the RMSD of ECR's prediction (ECR) and the solution number of the decoy (HEX Soln #); structure with the smallest RMSD within the FLIP cluster (ECR best RMSD).

**Table 4.2.** *Success of HEX Versus ECR*

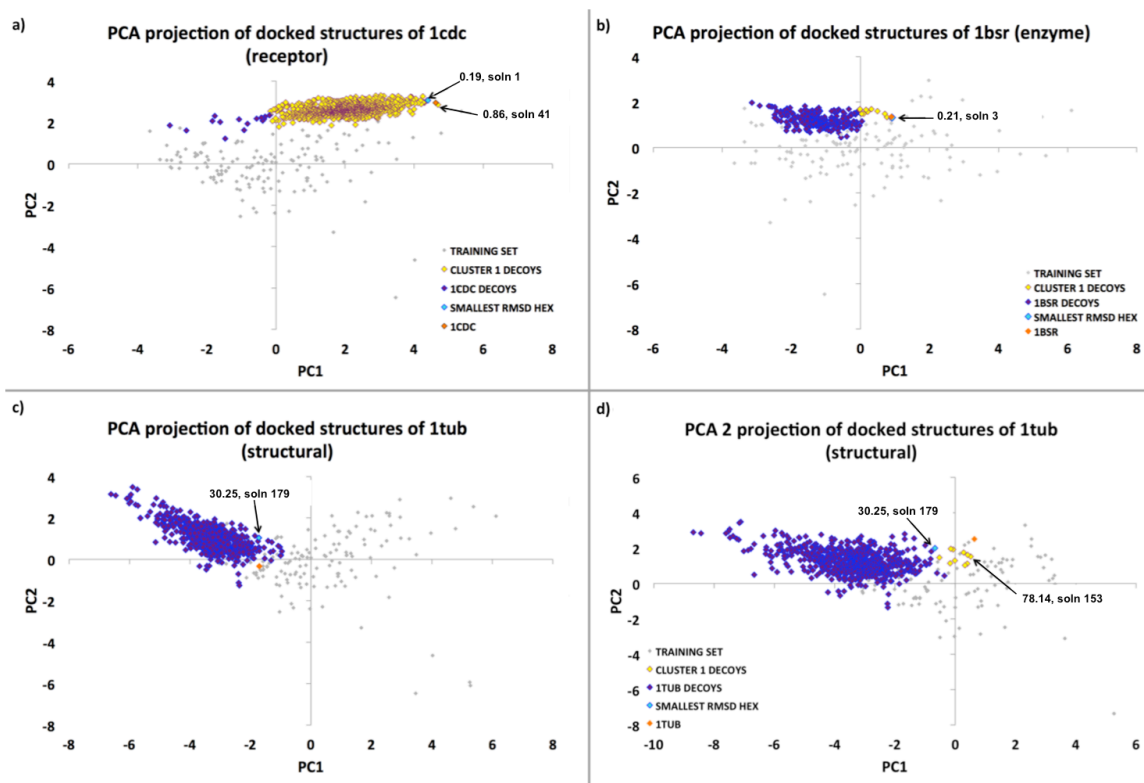
	<b>Success</b>	<b>AbAg</b>	<b>Inhibitor</b>
<b>1<sup>st</sup> Pose (HEX)</b>	9/22 (41%)	0/2	1/3
<b>1<sup>st</sup> Pose (ECR)</b>	11/22 (50%)	0/2	1/3
<b>1000 Poses (HEX)</b>	12/22 (54%)	0/2	2/3
<b>FLIP Cluster Poses (ECR)</b>	10/22 (45%)	0/2	1/3

The success of predictions in best 1 and 1000 poses for both HEX and ECR. No AbAg interfaces were correctly predicted either by HEX or ECR. In predicting inhibitors, HEX and ECR identified only 1 of 3 native-like interfaces (33%) when considering only the top pose. However, when all thousand poses were considered, HEX identified 2 interfaces (66%). When all poses from the FLIP cluster were considered, we achieved the same success rate.

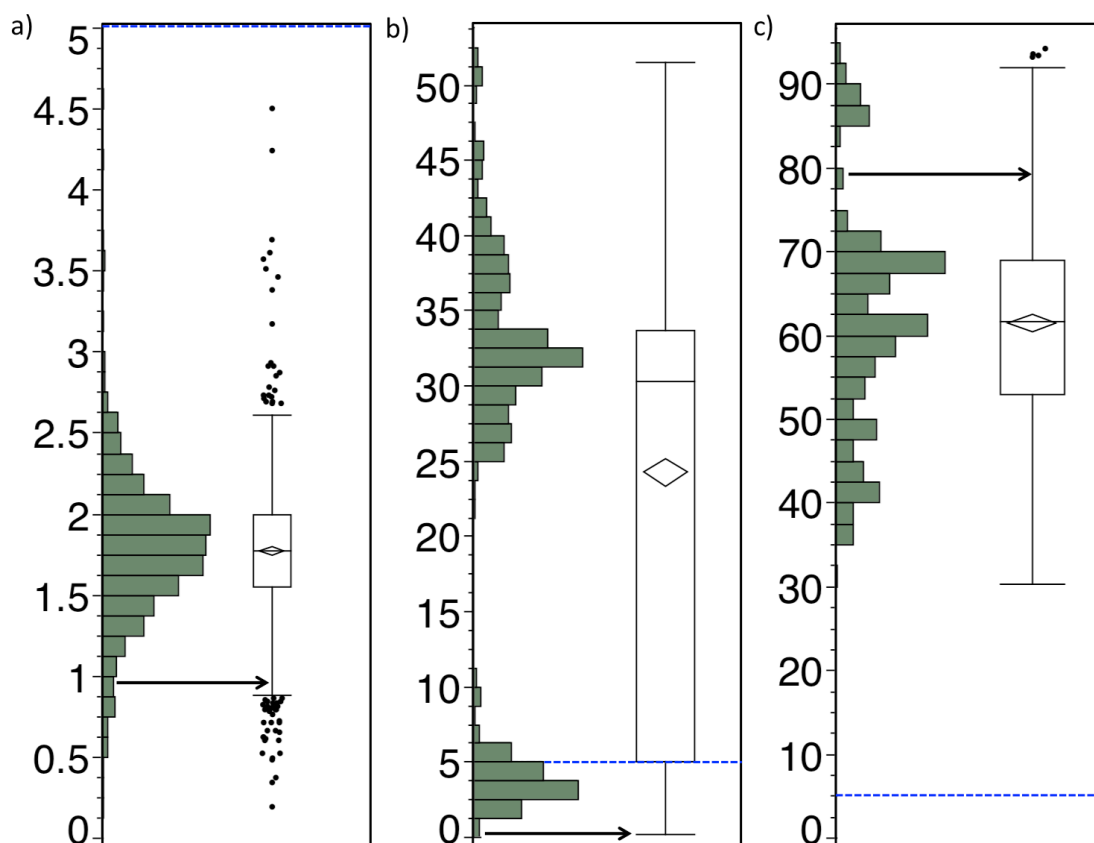
**Table 4.3.** *Analysis of Distribution Plots*

<b>Category</b>	<b># proteins</b>	<b># correct</b>	<b># incorrect</b>	<b># lower quartile</b>	<b># upper quartile</b>	<b># upper and incorrect</b>	<b># HEX failures</b>
<b>FLIP</b>	18	10/18	8/18	13/18	5/18	4/5	8/18
<b>FunC</b>	4	0/4	4/4	2/4	2/4	2/2	4/4

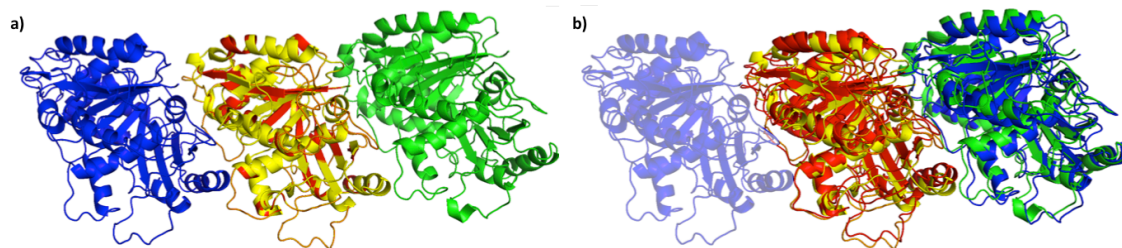
The count of correct and incorrect predictions of ECR for FLIPs and FunCs. The number of interfaces for which ECR's prediction was in the top or bottom quartile are indicated. The number of times ECR's prediction was in the upper quartile and was also the incorrect solution as well as the number of times HEX failed are shown.



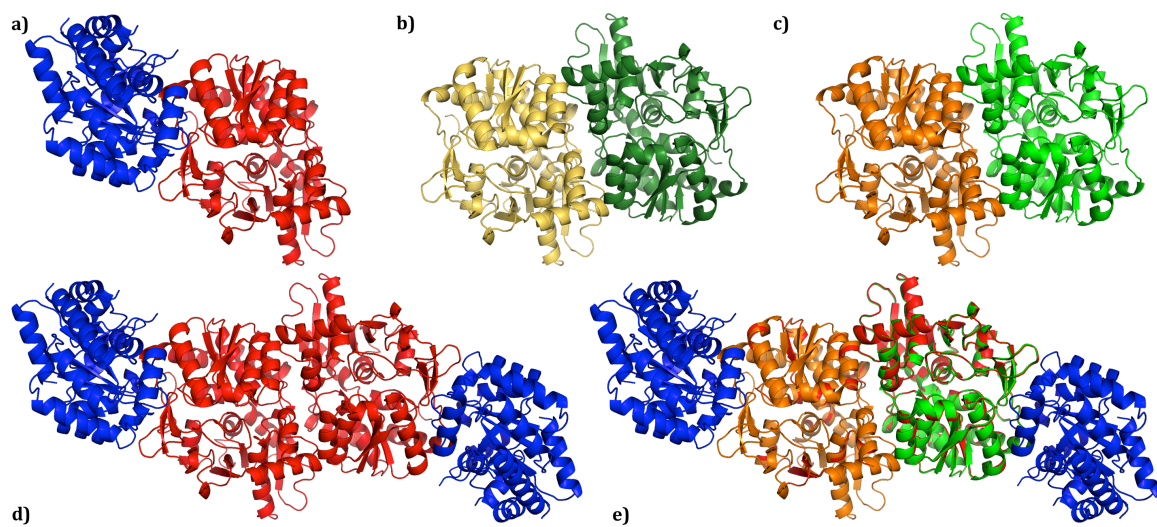
**Figure 4.1.** Distribution of docking poses of representative PPIs in FLIPlite. ECR analysis on PPIs partitions them into FLIP (yellow) and FunC (purple) clusters. ECR values for the known PDB structure for each example are indicated in orange. The structure with the smallest RMSD from the known PDB structure is indicated in cyan. HEX's number 1 solution is also indicated for each example. The ECR solution in all cases is the structure with the most positive PC1.  
a). *1cdc*, b). *1bsr*, c). *1tub*: 1<sup>st</sup> PCA projection, d). *1tub*: 2<sup>nd</sup> PCA projection.



**Figure 4.2.** Histogram and box plots of RMSDs of poses relative to the known structure. For each box plot, the ends of the box represent the interquartile range. A horizontal line inside the box marks the median. The top and bottom points of the diamond represent the upper and lower 95% from the mean. The whiskers are  $1.5 \times$  interquartile range from the first and third quartile. The dots on the outside of each whisker are outliers. The arrow indicates the location of ECR's chosen pose in the distribution and the blue line indicates the 5 Å cut-off, below which all poses are considered native-like. a) *Icdc*, b) *Ibsr*, and c) *Itub*

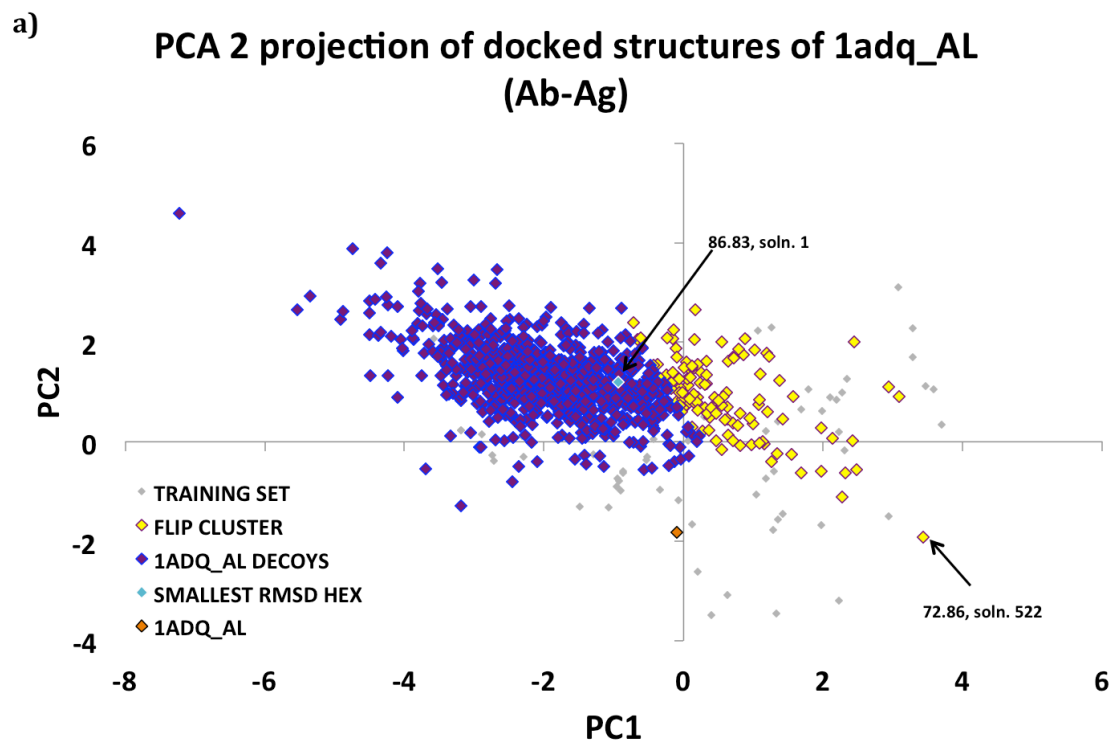


**Figure 4.3.** ECR prediction of *Itub*. a) The PDB structure of *Itub*, chains A (red) and B (blue). Also seen in a) is the HEX pose identified by ECR - solution 153 (chain A: yellow and chain B: green). The RMSD on super-positioning these structures is 78.14. b) When the same solution (153) was superposed on the crystal symmetry transformed conformations of *Itub*, chains A (red) and B (blue), an RMSD of 2.2Å was obtained.

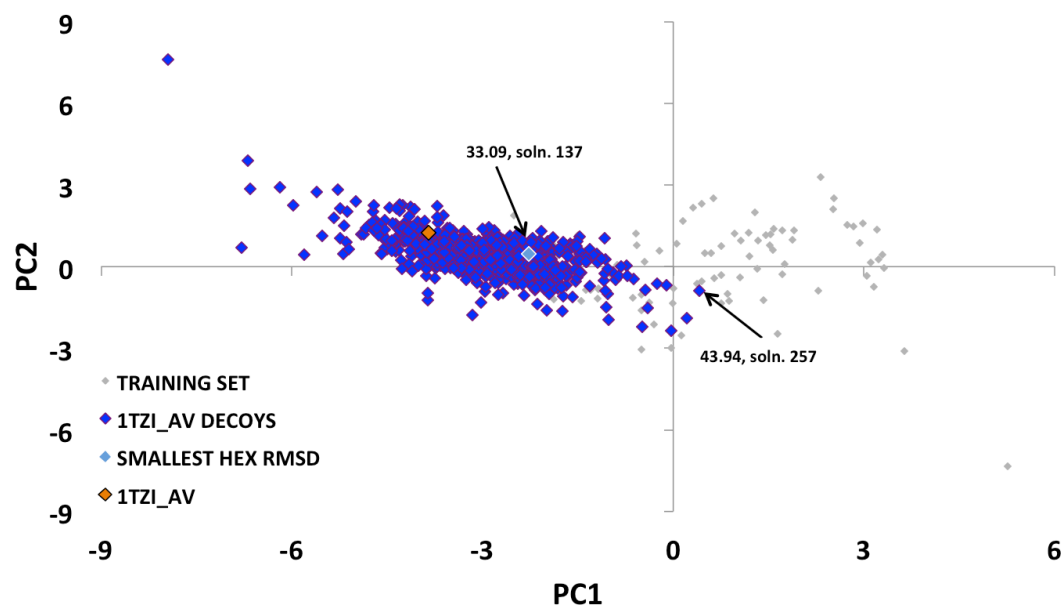


**Figure 4.4.** ECR prediction of *xIubs1*. (a) The PDB structure of *Iubs* chains A in blue and B in red. (d) The functional tetramer of *Iubs* (dimer of dimers). (b) The structure of *xIubs1* chains B in yellow and Z (a copy of chain B) in dark green. (c) ECR's pick, solution 2 from docking *xIubs1* with chains B in orange and Z in green. (e) A superposition of *xIubs1* and ECR's pick had an RMSD of 0.33Å.

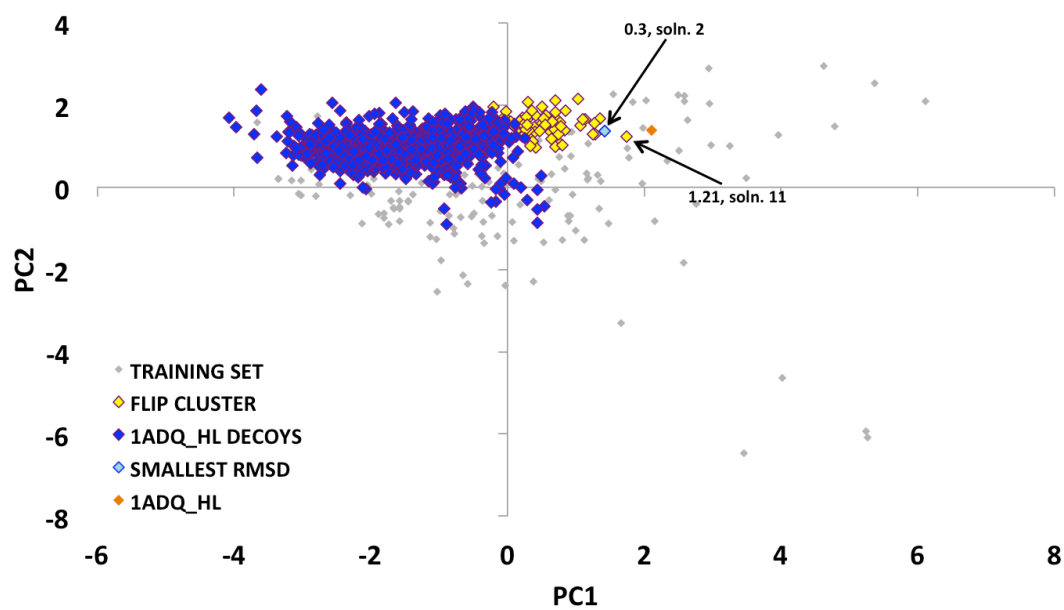
## Supplementary Figures

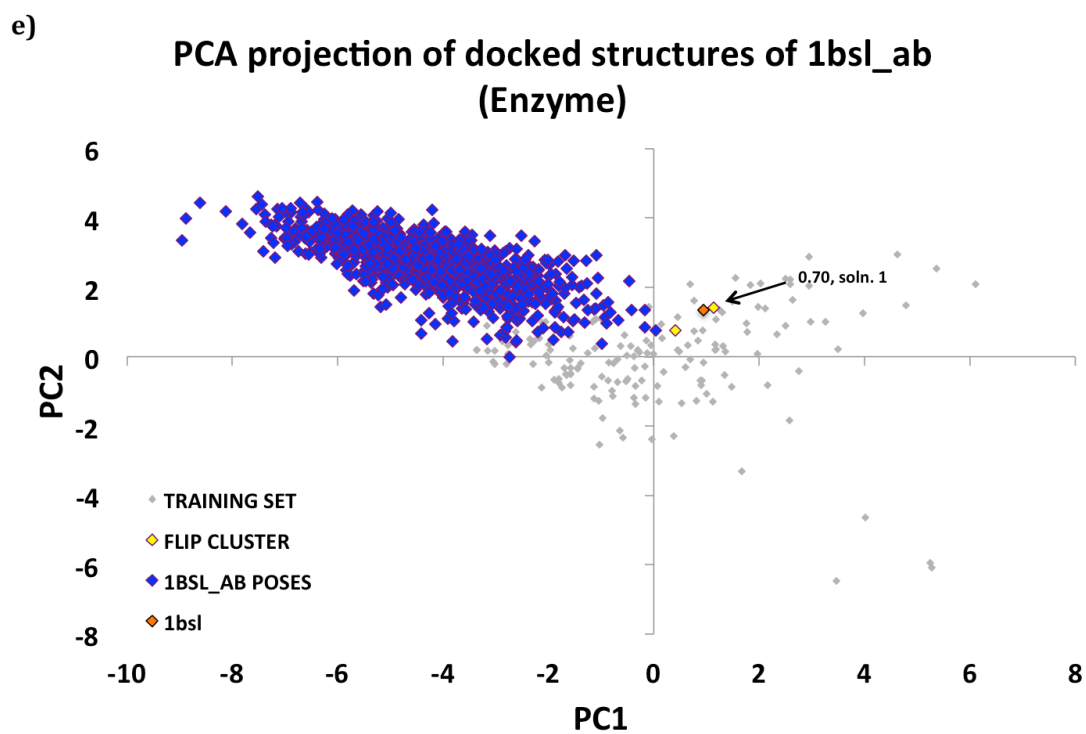
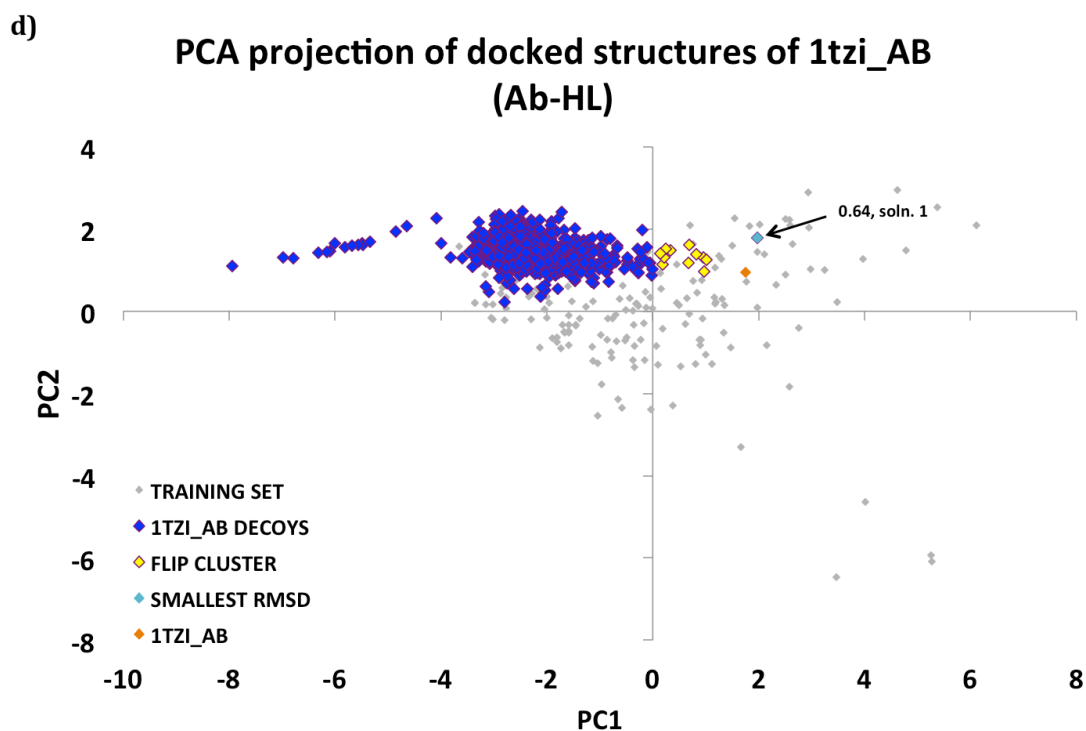


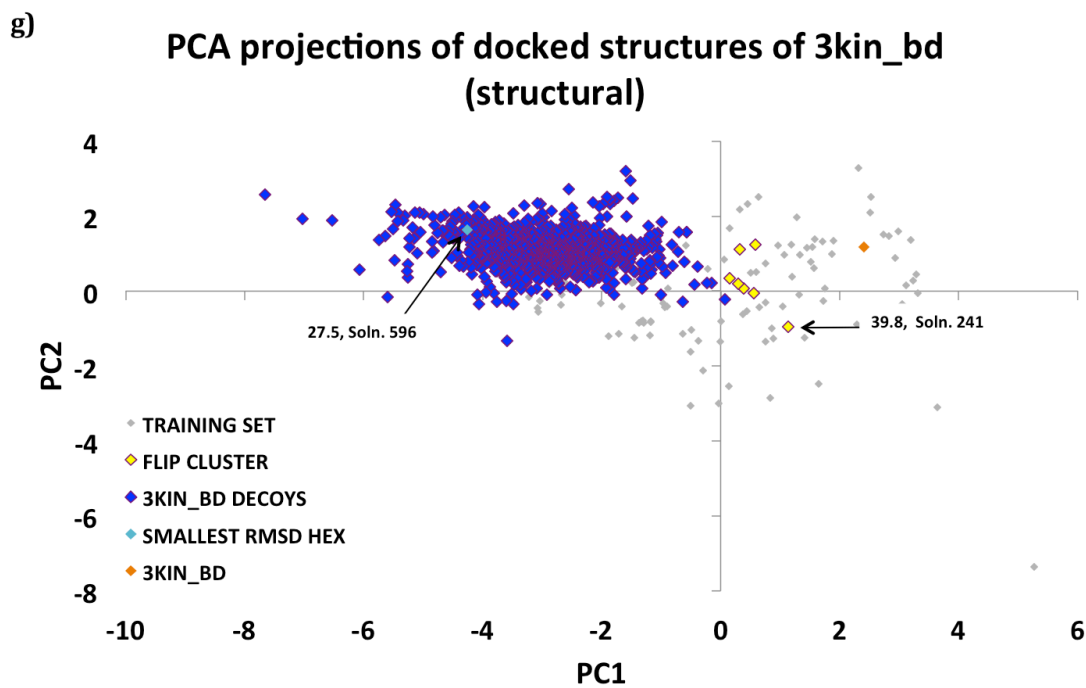
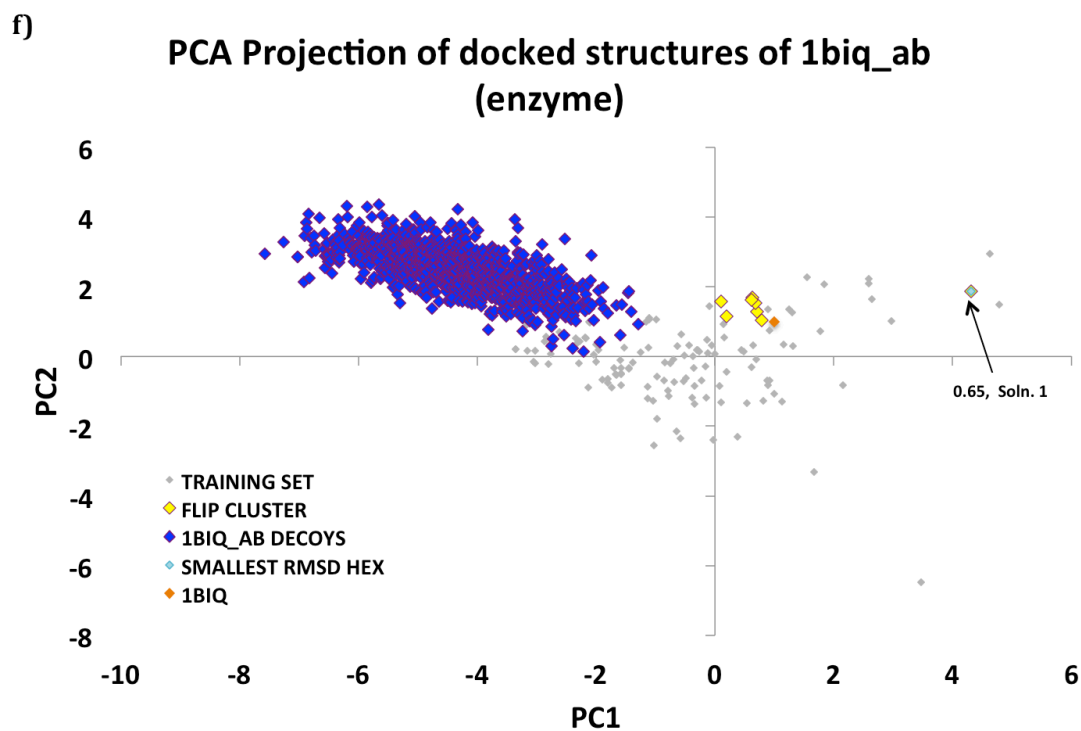
b) **PCA 2 projection of docked structures of 1tzi\_AV  
(Ab-Ag)**



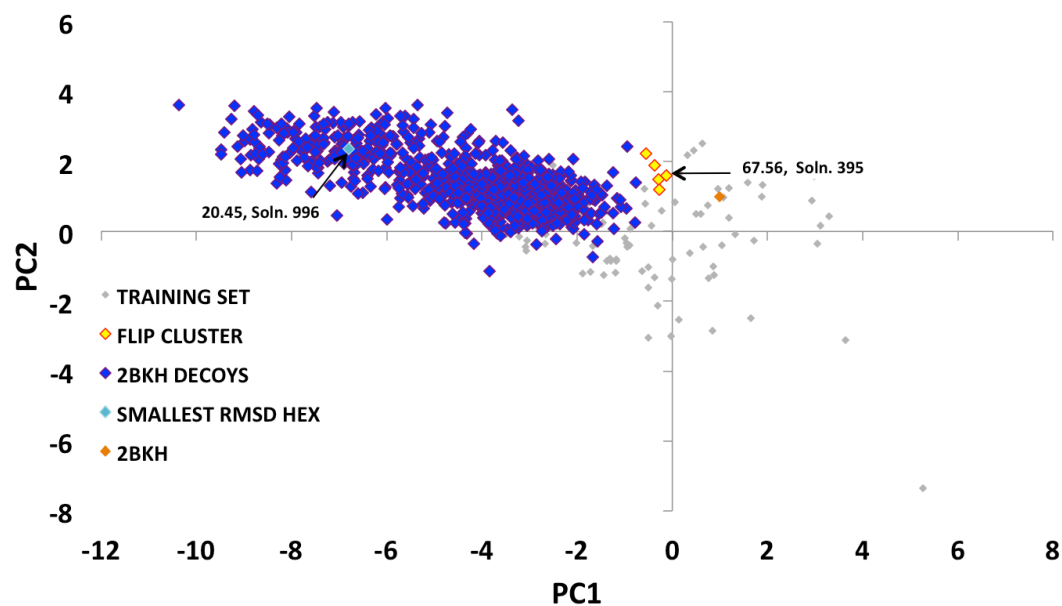
c) **PCA projection of docked structures of 1adq\_HL  
(Ab-HL)**



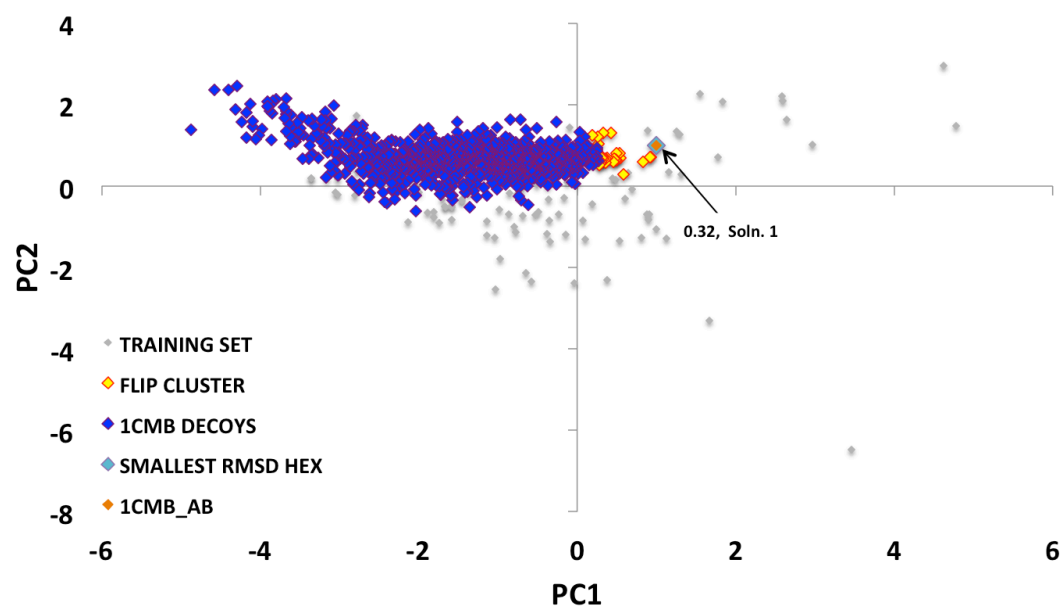


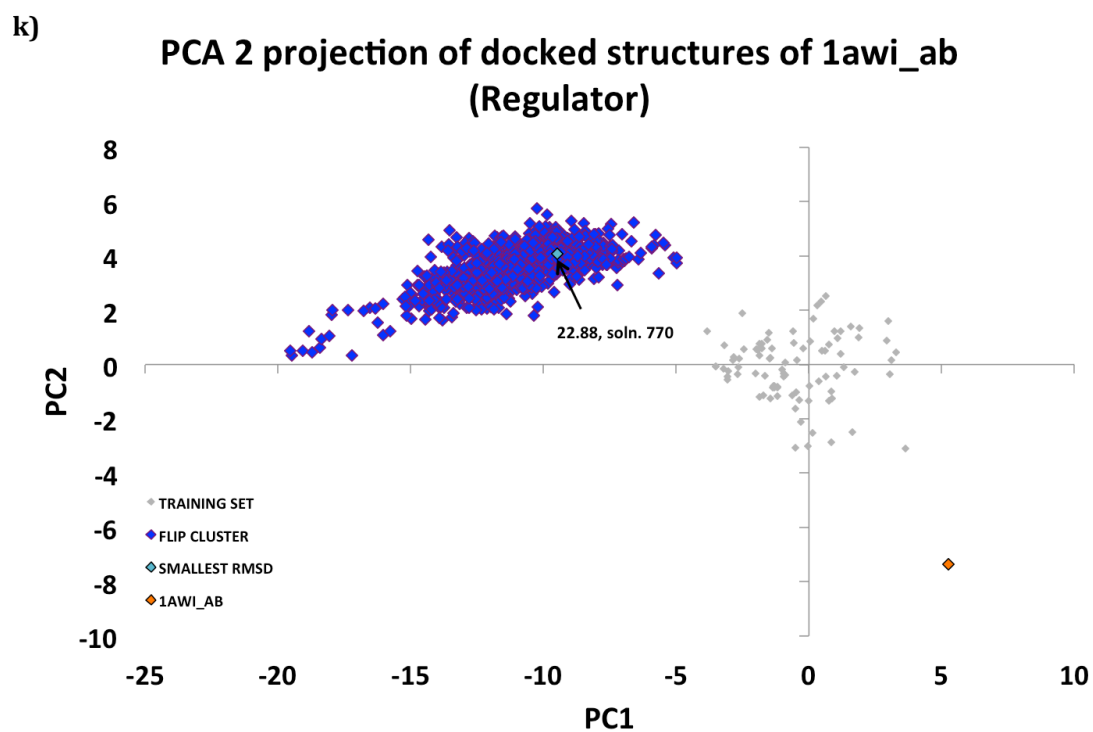
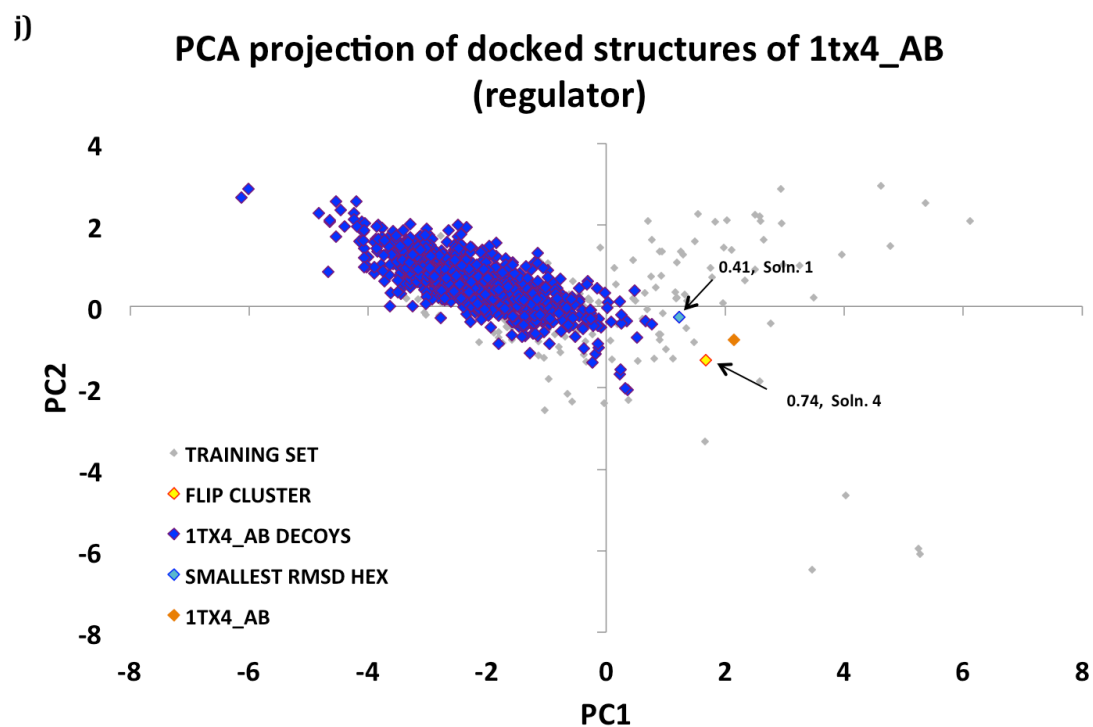


h) **PCA 2 projection of docked structures of 2bkh\_ab (structural)**



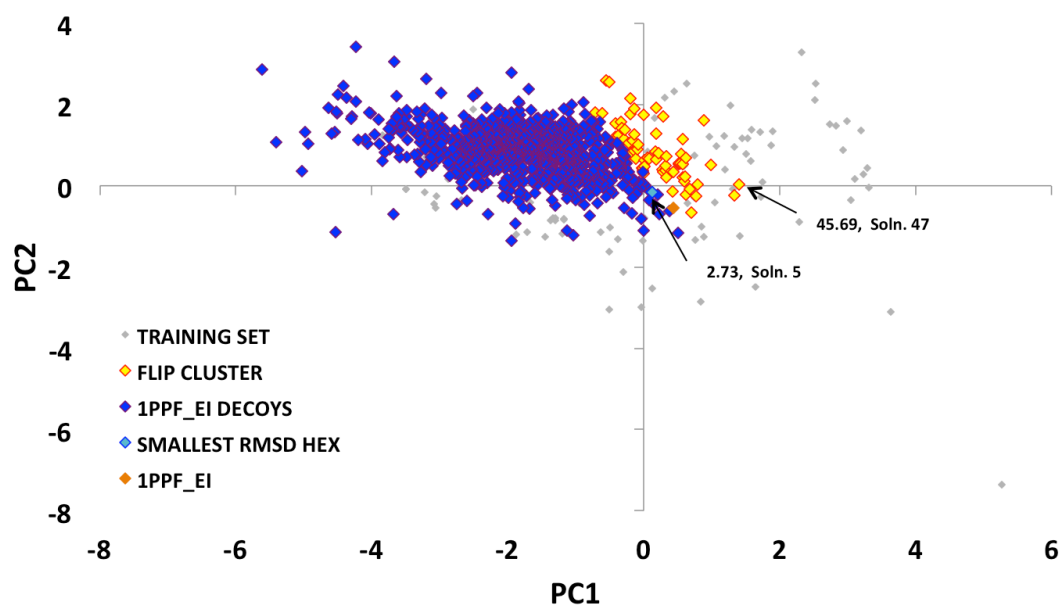
i) **PCA projections of docked structures of 1cmb\_ab (regulator)**





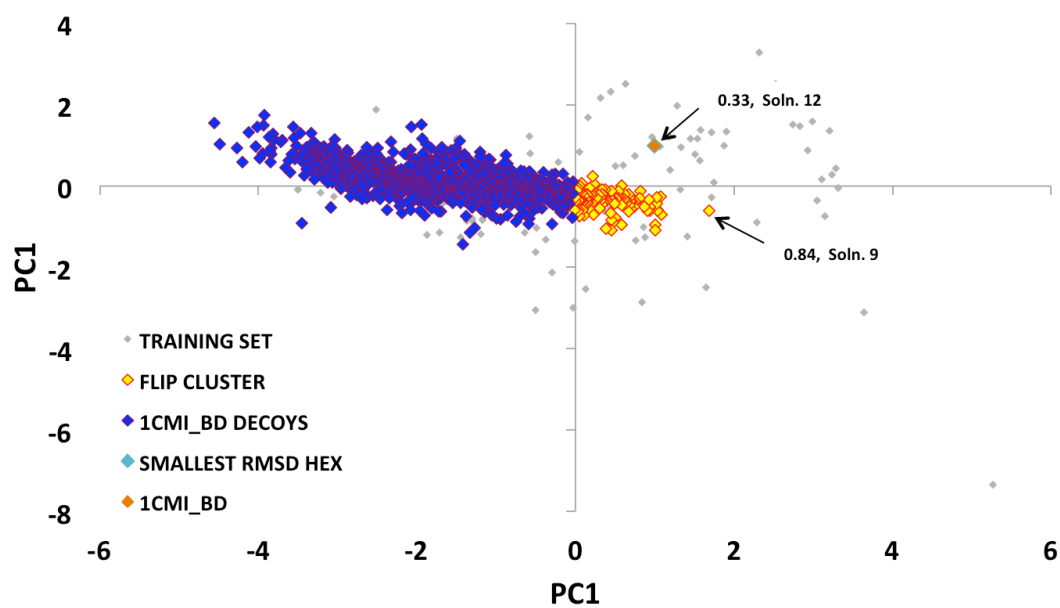
l)

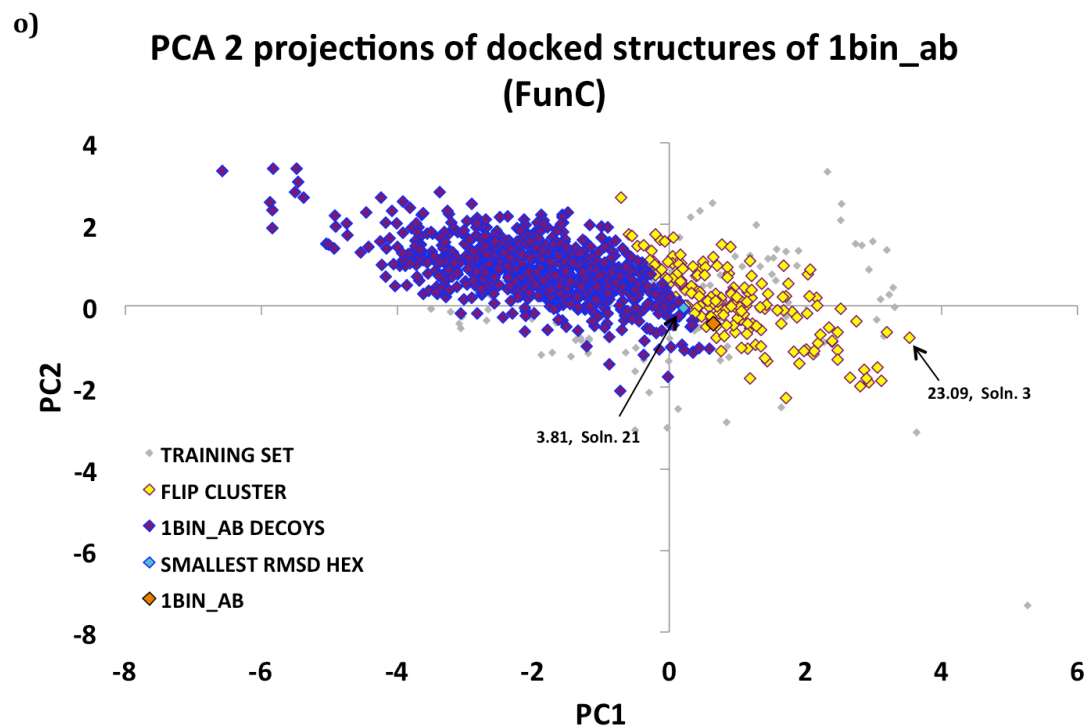
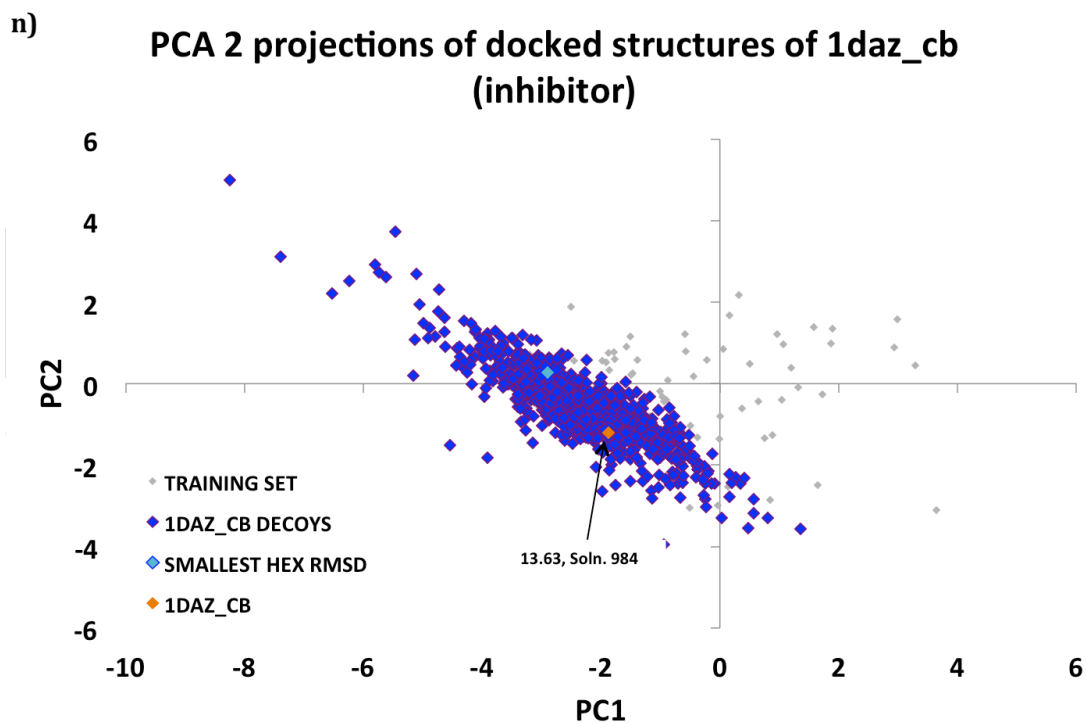
### PCA projections of docked structures of 1ppf\_ei (inhibitor)



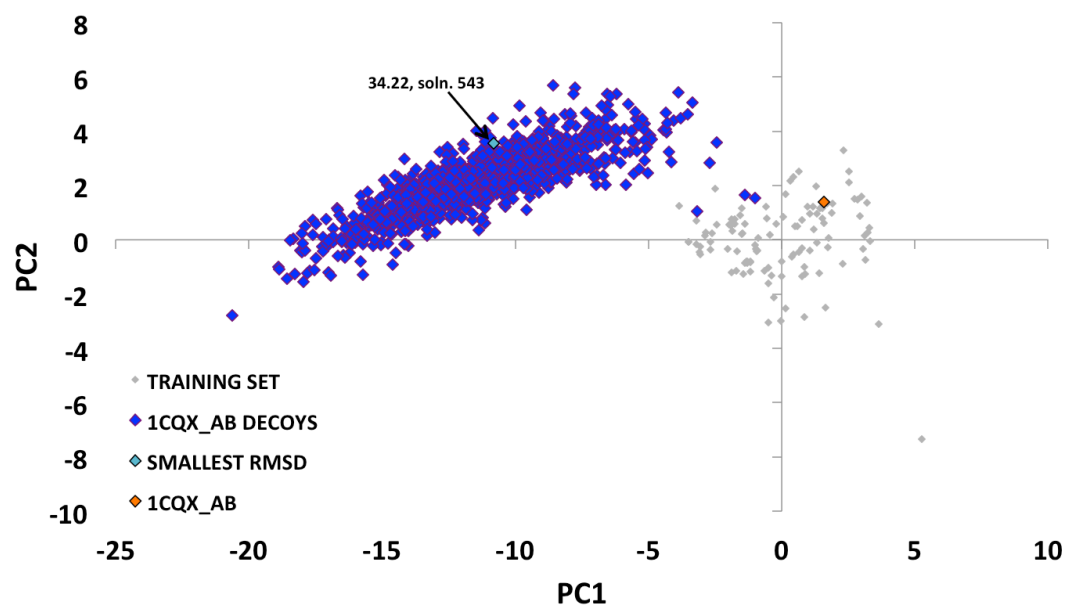
m)

### PCA 2 projections of docked decoys of 1cmi\_bd (inhibitor)

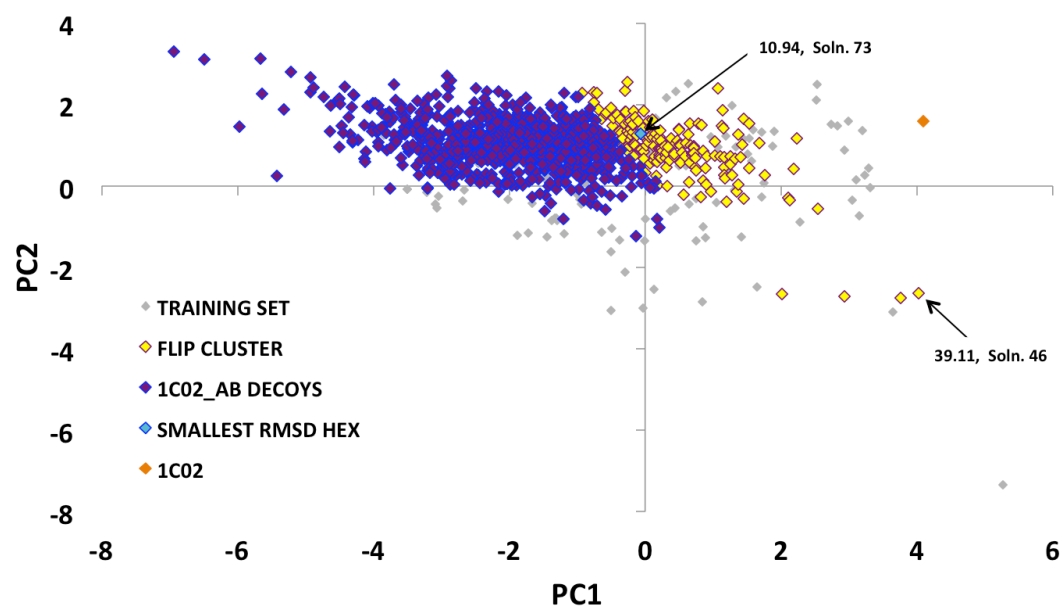




p) **PCA 2 Projection of docked structures of 1cq<sub>x</sub>\_ab (FunC)**

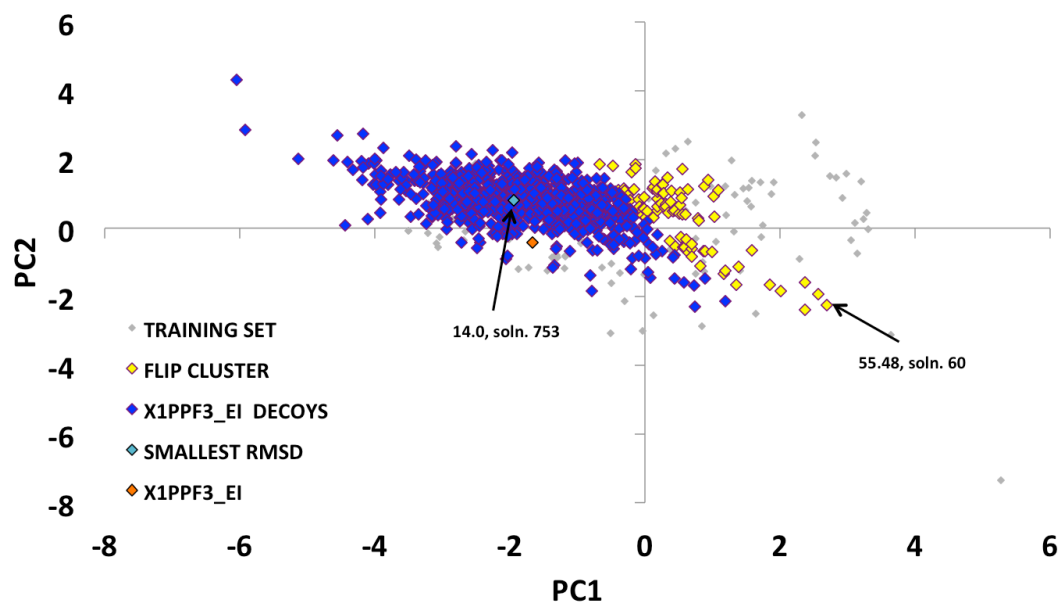


q) **PCA 2 projections of docked structures of 1c02\_ab (FunC)**



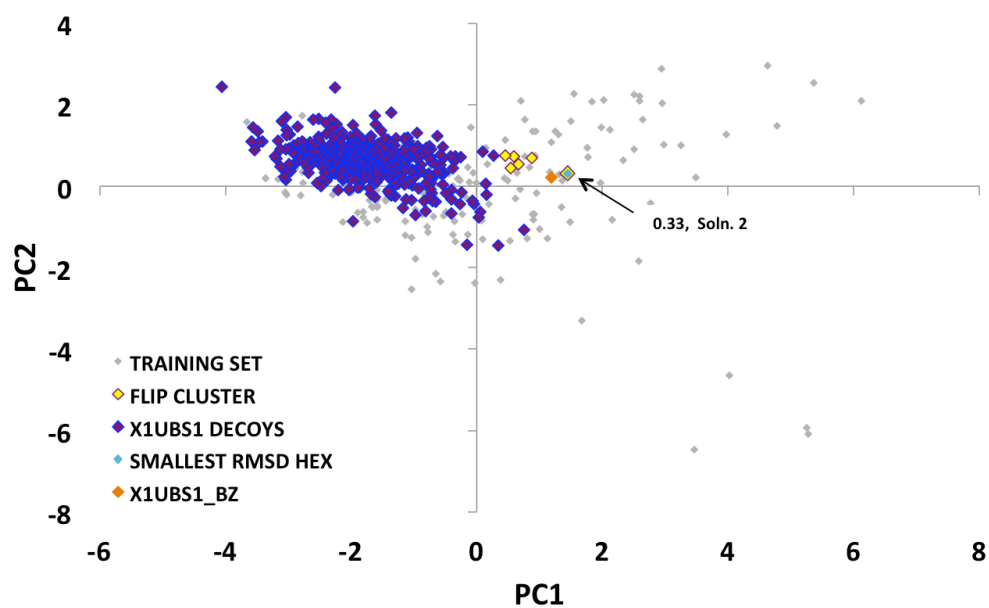
r)

### PCA 2 projection of docked structures of x1ppf3\_ei (XFunC)



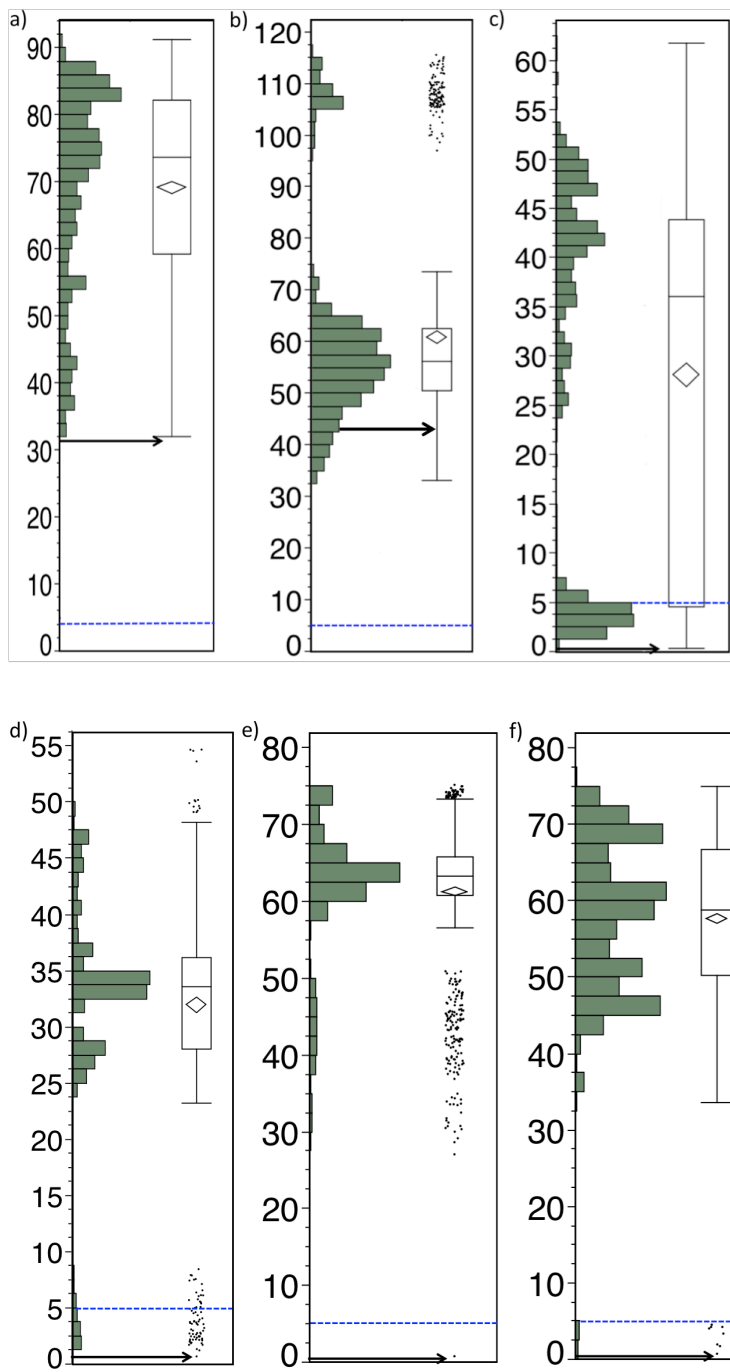
s)

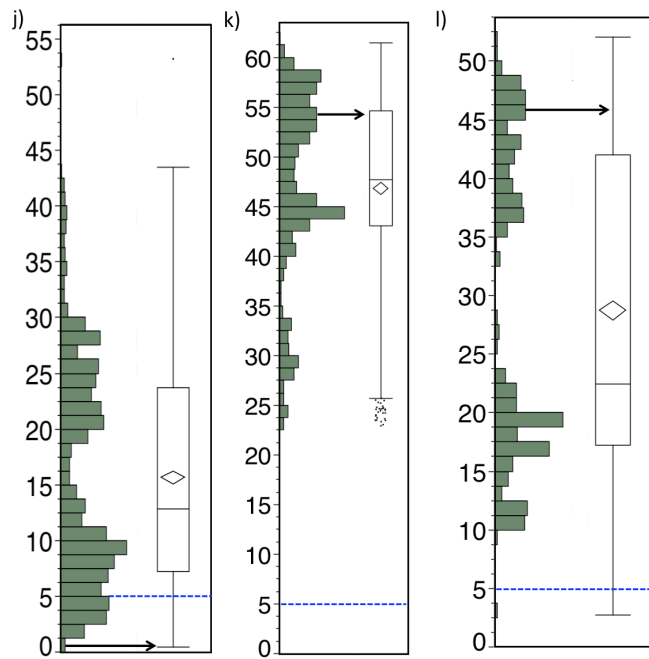
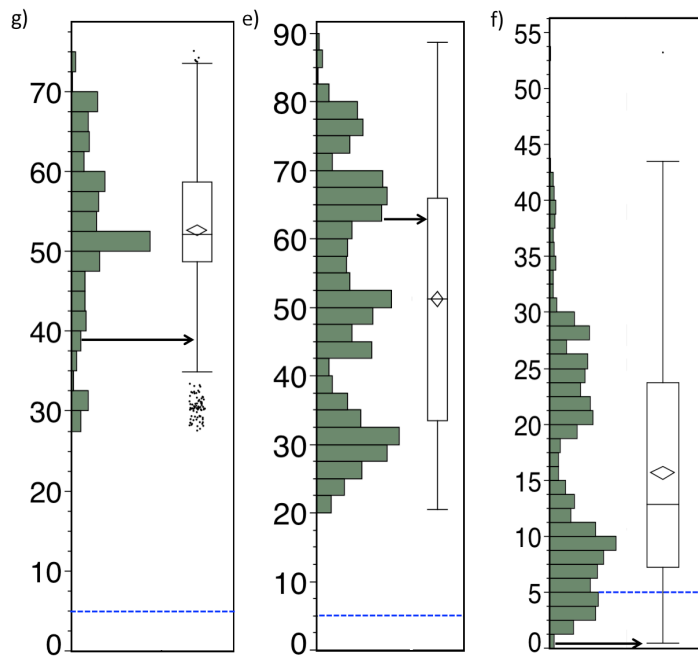
### PCA projections of docked structures of x1ubs1\_bz (xFunC)

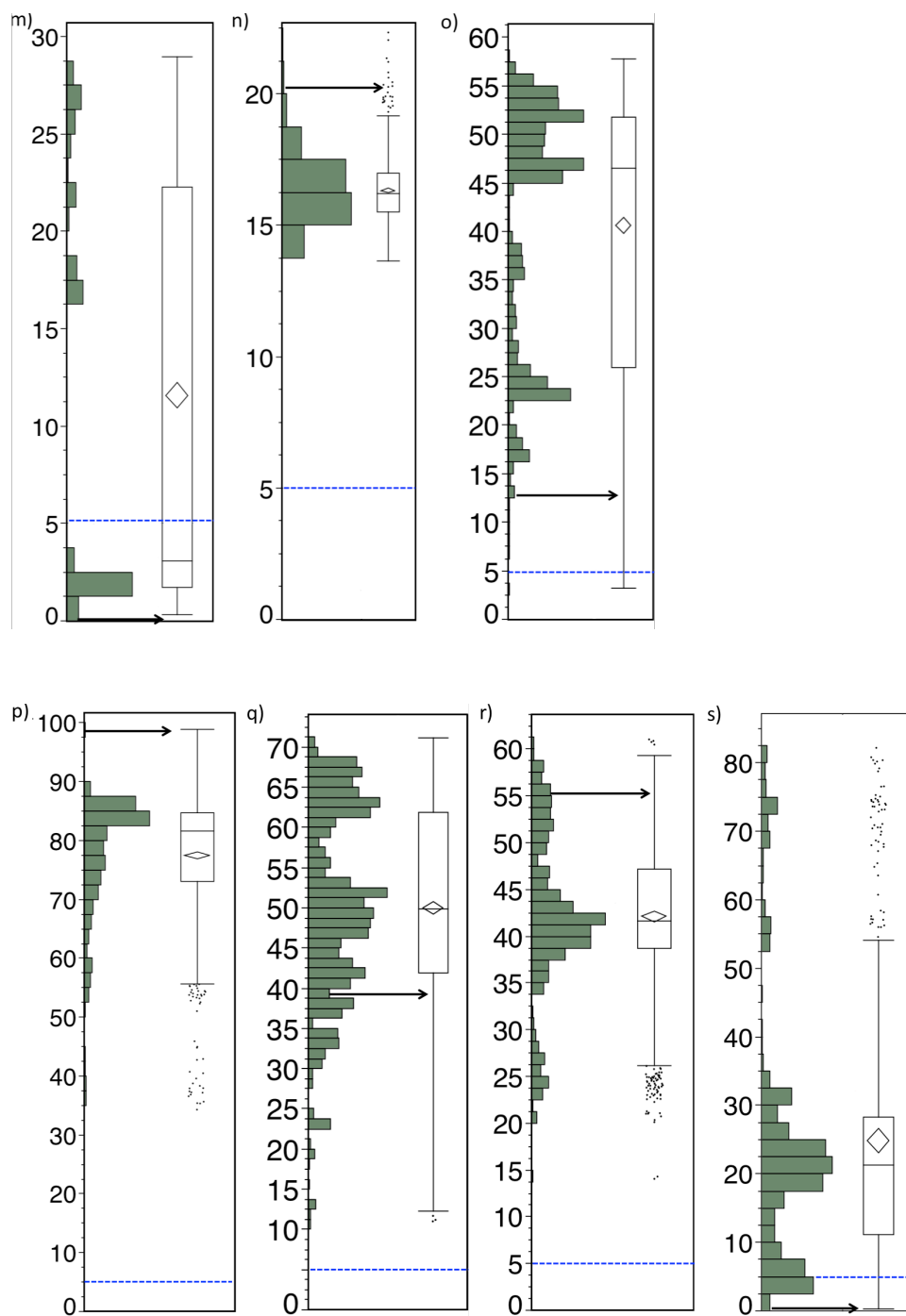


**Figure S4.1.** Distribution of docking poses of PPIs in FLIPlite. ECR analysis on PPIs partitions them into FLIP (yellow) and FunC (purple) clusters. ECR values for the known PDB structure for each example are indicated in orange. The structure with the smallest RMSD from the known PDB structure is indicated in cyan. HEX's number 1 solution is also indicated for each example. The ECR solution in all cases is the structure with the most positive PC1.

a. 1adq\_al; b. 1tzi\_av; c. 1adq\_hl;  
d. 1tzi\_ab; e. 1bsl\_ab; f. 1biq\_ab;  
g. 3kin\_bd; h. 2bkh\_ab; i. 1cmb\_ab;  
j. 1tx4\_ab; k. 1awi\_ab; l. 1ppf\_ei;  
m. 1cmi\_bd; n. 1daz\_cb; o. 1bin\_ab;  
p. 1cqx\_ab; q. 1c02\_ab; r. x1ppf3\_ei;  
s. x1ubs1\_bz







**Figure S4.2.** Histogram and box plots of RMSDs of poses relative to the known structure. For each box plot, the ends of the box represent the interquartile range. A horizontal line inside the box marks the median. The top and bottom points of the diamond represent the upper and lower 95% from the mean. The whiskers are 1.5 \* interquartile range from the first and third quartile. The dots on the outside of each whisker are outliers. The arrow indicates the location of ECR's chosen pose in the distribution and the blue line indicates the 5 Å cut-off, below which all poses are considered native-like.

a. 1adq\_al; b. 1tzi\_av; c. 1adq\_hl;  
d. 1tzi\_ab; e. 1bsl\_ab; f. 1biq\_ab;  
g. 3kin\_bd; h. 2bkh\_ab; i. 1cmb\_ab;  
j. 1tx4\_ab; k. 1awi\_ab; l. 1ppf\_ei;  
m. 1cmi\_bd; n. 1daz\_cb; o. 1bin\_ab;  
p. 1cqx\_ab; q. 1c02\_ab; r. x1ppf3\_ei;  
s. x1ubsl\_bz

## CHAPTER V

### CONCLUSION

One of the goals of molecular and cellular biology is identifying functions of various molecules. Function is dependent on protein quaternary structure, which makes the prediction of binding modes important. However, binding is still not entirely understood. Since it has been established that only a few amino acids are critical to maintaining structure, function or both, the identification of such residues may improve prediction of quaternary interaction. This in turn could improve functional annotation of protein structures of unknown function and thus improve docking and drug design studies. Various approaches focus on developing scoring functions to distinguish interface from non-interface using features such as hydrophobicity, sequence conservation, solvent accessible surface area, and shape complementarity. These approaches have yielded limited to moderate success<sup>1, 2</sup>. Similar success has been seen with machine learning approaches like support vector machine (SVM) and neural networks that use sequence and structure information as input to predict quaternary structures<sup>3</sup>.

In this work, our aim was to identify properties of PPIs that could discriminate between operationally defined FLIP and FunC categories of interfaces using both interface properties and machine learning methods. We constructed a dataset of 160

interfaces, categorized by function, called FLIPdb. We also introduced the Energy Centrality Relationship (ECR) concept. Simplistically, ECR presents as a central organizing tendency, with more hot spots at the center of the interface and residues with radially diminishing energetic importance around the center. Even though it is known that residues at the interface contribute variably<sup>4</sup> to its stability, it is generally also known that interfaces consist of subsets of residues that contribute to the bulk of the total binding energy<sup>5</sup>. This is of particular importance in drug designing, as the residues contributing to a high cumulative binding energy at the interface can be prime targets for drug binding. Despite solely using energetic and geometric features, ECR has a success of 76-80%, a distinct advantage over methods that use sequence conservation and interactome maps because energy is directly correlated to function and stability while conservation is not. Although ECR's success is somewhat lower than some of the best predictors, it does not appear to have major database compositional bias. From the analysis of computational alanine scanning mutagenesis (CAS), free-energy of substitution ( $\Delta\Delta G$ ) and geometric data, FLIPs demonstrate more specific amino acid interactions that produce larger disruptions on alanine substitutions than those of FunCs. This finding is consistent with experimental work<sup>6</sup> and with the characterization of specific and non-specific interfaces<sup>7</sup>.

The occurrence of both a central organizing tendency and a larger specificity at the interface suggest a model for evolution at the interface, in which interfaces start as weak associations and grow over evolutionary time under a natural selective pressure to maintain or even “improved” (via mutation) the proteins' biological function(s)

(Figure 2.2c-f). Residues surrounding these contacts may also have pressure to improve affinity. The result of these selective pressures over evolutionary time on the size and specific affinity produce a radially symmetric pattern (Figure 2.2b,f) demonstrating “stronger” energies near the “older” regions of the interface. This hypothesis agrees with the Evolutionary Trace results of Lichtarge and colleagues, who identify radially symmetric “bulls-eye” sequence conservation patterns near functionally important residues<sup>8</sup>.

It has been observed that the evolutionary rate of proteins is constrained by factors such as structural and functional characteristics, pressures proportional to mRNA expression levels, and the role of protein abundance. At a higher population size and in the absence of genetic drift, proteins are intrinsically stable<sup>9</sup>. Proteins that fail to evolve intrinsic stability are generally stabilized by PPI. The existence of a trade-off between the requirement for maintenance of sufficient stability and the necessity to avoid aggregation is also known<sup>10</sup>. Aggregation prone regions are flanked by “gatekeepers” residues that are conserved even at a high thermodynamic cost to the protein<sup>11</sup>. This suggests that a selective pressure is placed on the protein to interact with a particular partner with high affinity suggesting these residues are not structurally, but evolutionarily limited, in order to prevent protein aggregation. This requirement for biochemical conservation may be what protects an interface from genetic drift.

Our conservation score analysis showed limited trends. A comparison of evolutionary data with interface geometries during analysis of the training set seem to suggest FLIP interfaces are likely to have different residue characteristics than FunC

interfaces. Consistent with our hypothesis of FLIPs having a central tendency, FLIPs have a positive correlation with the regression slope of the alignment variability meaning less variation occurs near the interface center relative to the edges. FunCs, on the other hand, seem to be much more variable in conservation patterns and indeed are generally negatively correlated with slope, meaning they have more variation towards the center. However, contrary to our expectations, FLIPs had a larger overall variability at the interface suggesting FLIP interfacial residues may demonstrate larger sequence variation during the co-evolutionary “optimization”, as multiple mutations may improve energy at the expense of increased variation. This is also consistent with the symmetric “bulls-eye” pattern of sequence conservation near functionally important residues as observed by Lichtarge and colleagues in numerous systems<sup>8</sup>.

Our hybridization of energetic analysis coupled with sequence conservation studies was not an improvement over our results from energy or sequence traits alone, which was also contrary to our expectations. This could be due to the co-evolutionary optimization events mentioned before. Natural selection of an interface could allow for variant interacting residue pairs while maintaining stability. This could have the effect of increased sequence variation, thus disrupting its correlation with geometric patterns.

A comparison of the accuracy of prediction for both energy and sequence conservation based analysis showed that at 76%, energetic analysis gave us superior results compared to sequence conservation, which had a success of 69%. Unlike with energetic analysis, our analysis of sequence conservation seemed to be somewhat compositionally biased. Therefore, the relationship between sequence conservation and

geometry, as we have used here, does not seem to be as reliable a FLIP/FunC discriminator as energetic criteria.

Detecting interacting regions and interacting partners of a protein helps in understanding large amounts of sequence and structural data. A combination of experimental and computational methods and analysis of the large amount of data accumulated from these sources would not only help in understanding how various proteins function within the cell, but would also lead to better ways of manipulating interfaces such as with drugs made for that purpose. Discovering a new drug is computationally expensive and is a process that generally takes decades. By narrowing the number of compounds that might treat a particular disease or by assembling novel drug molecules to disrupt specific disease pathways, computational methods could not only reduce the research time line but also the cost of experimentation. Some of the best-known examples of computer aided drug design are the inhibitors of the HIV protease and the HIV life cycle<sup>12</sup>. Our goal was to apply the ECR methodology to molecular docking runs of proteins to filter the number of possible solutions to a few (or one) best candidates.

Molecular docking is a conformational sampling method where receptor and ligand interact to generate many possible interaction conformations. The low energy conformations are then selected as the possible native state solutions. Docking should be guided by a scoring function but current scoring functions also generate a high number of false positive poses<sup>13</sup>. Re-sampling conformational space could therefore provide a pathway, much like the protein folding free-energy landscape, to generate conformations

with improved biochemical affinity. Such conformational targeting during binding may effectively provide a better path for association prediction.

Since ECR had ~80% success in predicting FLIPs and FunCs, we incorporated it as a post-filter to our docking analysis. The docked poses of structures in FLIPdb-lite were projected through our PCA analysis of the training set. The poses generated for most structures seemed to mostly occupy quadrant 2 (FunC region) with some poses in quadrant 1 (FLIP region). The difference in the general distribution pattern of FLIPs, FunCs, and docking poses suggests a fundamental difference in the biology of these structures or interfaces. The poses generated by HEX are the top predictions for the docker and are not random predictions. Still, the generation of a high number of false positives indicates that HEX possibly focuses on predicting complexes with high affinity, which is important to the stability of the complex. However, high affinity does not necessarily mean high specificity. Distinguishing affinity from specificity may be essential for fine-tuning molecular interaction predictors. Using ECR as a post-filter, we were able to improve the success of prediction from 41% to 50%. While this may not be a dramatic improvement over HEX, it shows us that ECR identifies features either ignored by or weighed less by HEX. One of the major contributions of ECR as a post-filter is the ability to identify alternate FLIP interfaces. Identification of the correct interface to disrupt is of particular importance in drug-design.

Apart from using the appropriate scoring function to identify native-like structures, one other challenge is the incorporation of flexibility. The proteins studied in

our CAPRI-like experiment in Appendix 1, unlike the previous docking experiments in chapter 3, involved at least one unbound component. The difficulty of unbound docking is higher as it has to account for both backbone and side-chain conformational changes that occur upon binding. For such interactions, if a docking program does not incorporate flexibility into the calculations, the probability of generating false-positives is high. As ECR is based on studying static structures it may not be an effective post-filter in such cases. This might be one reason why ECR failed to identify native-like poses in all three targets selected. Calculations are currently underway that address ECR's sensitivity to conformational changes (unpublished results).

Furthermore, ECR consistently clustered AbAg and Enzyme-Inhibitor interfaces near FunCs and away from other FLIPs using either residue energetics or conservation features. As a result, our docking analysis of these sub-categories was inconclusive. Even though the distribution of interfaces belonging to these sub-categories were internally consistent, in that they clustered together, they generally were in the FunC region. It is possible that these sub-categories were not properly operationally defined to start with or that we did not identify the features important to these categories. For example, the antigen binding site on an antibody has loop conformations called Complementarity Determining Regions (CDRs). The conformations of these loops are determined by a few key residues within each loop or in interacting loops. These residues are conserved while the surrounding residues tend to vary<sup>14</sup>. With this type of multiple instance of conserved and variable regions, it may be that AbAg appear to lack overall geometric centrality when in actuality they have multiple tiny 'central' regions.

AbAg interactions are not driven by natural selection of mutants over evolutionary time periods but rather a result of stochastic V(D)J recombination within the lifespan of the organism<sup>15, 16</sup>. Similarly, infectious organisms often produce enzyme inhibitors to hamper a host's native functions. While, the infecting organism may have a selective pressure to enhance inhibitor binding, the host organism actually has selective pressures to evade inhibitor binding. These alternative pressures on such interfaces may also produce more FunC-like geometric patterns.

The size of the interface itself could be a factor in FLIP/FunC prediction of AbAg and Enzyme-Inhibitor complexes. Smaller interface sizes would mean fewer residues and overall lower correlation values for the features chosen for our K-means clustering analysis, which is probably another reason these sub-categories were difficult to distinguish from the FunCs.

The CAPRI-like docking analysis included two Enzyme-Inhibitor complexes. Our low success with this sub-category, coupled with the proteins not being in their final bound conformations, likely made it more difficult to predict them. Further analysis in this area could be to identify unbound components of the proteins in FLIPlite (which is dominated by bound PPI) and perform docking studies on them. A comparison could then be made between the two sets to see if the lack of success in our CAPRI-like docking analysis was due to issues with improper definitions of sub-categories or flexibility within the interface.

ECR currently is based on analyzing coordinate files from PDB representing static structures. Proteins however are dynamic and HEX failed to generate native-like

solutions for all three CAPRI targets chosen. This could either be because of protein flexibility and dynamics or just because it is a challenging sub-category. While we could use docking algorithms that are better suited for certain proteins, for example GLIDE<sup>17, 18</sup>, a flexible docking algorithm for proteins that undergo large conformational changes upon binding; Rosetta Antibody<sup>19</sup>, for antibody interactions; and autodock<sup>20</sup>, for small ligand interactions such as enzyme-inhibitor interactions. We could also simulate backbone and side-chain movements using molecular dynamics simulation experiments to generate additional native-like low energy conformations that can be used as starting conformations for docking.

In summary, our ECR model shows the importance of energy in interface formation. It also shows that, even though sequence conservation is observed in proteins, it may not be sufficient for a FLIP/FunC discrimination. In combination, the use of energy features outperforms the use of sequence conservation features. Using ECR as a post-filter to docking reduces the number of false positives in bound docking. In studies with unbound components however, the benefit of using ECR was unclear. The inclusion of more molecular modeling methods during the modeling of the interfaces may improve predictions.

## References

1. Levy ED (2007) PiQSi: protein quaternary structure investigation. *Structure* 15:1364-1367.
2. Zacharias M (2005) ATTRACT: protein-protein docking in CAPRI using a reduced protein model. *Proteins* 60:252-256.
3. Zhu H, Domingues FS, Sommer I, Lengauer T (2006) NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics* 7:27-27.
4. Wells CT (1995) A hot spot of binding energy in a hormone-receptor interface. *Science* 267:383-386.
5. Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. *J Mol Biol* 280:1-9.
6. Bradshaw RT, Patel BH, Tate EW, Leatherbarrow RJ, Gould IR (2011) Comparing experimental and computational alanine scanning techniques for probing a prototypical protein-protein interaction. *Protein Eng Des Sel* 24:197-207.
7. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2004) A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol* 336:943-955.
8. Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 257:342-358.
9. Ohta T (1976) Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theoretical Population Biology* JID - 0256422.
10. Bastolla U, Moya A, Viguera E, van Ham RC (2004) Genomic determinants of protein folding thermodynamics in prokaryotic organisms. *Journal of Molecular Biology* JID - 2985088R.
11. Monsellier E, Chiti F (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Reports* JID - 100963049.
12. Olson AJ, Goodsell DS (1998) Automated docking and the search for HIV protease inhibitors. *SAR and QSAR in Environmental Research* JID - 9440156.
13. Perola E (2006) Minimizing false positives in kinase virtual screens. *Proteins* JID - 8700181.

14. MacCallum RM, Martin ACR, Thornton JM (1996) Antibody-antigen Interactions: Contact Analysis and Binding Site Topography. *J Mol Biol* 262:732-745.
15. Greenbaum JA, Andersen P, Blythe M, Bui H, Cachau R, Crowe J, Davies M, Kolaskar A, Lund O, Morrison S, et al (2007) Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *Journal of Molecular Recognition : JMR JID* - 9004580.
16. Kuroda D, Shirai H, Jacobson MP, Nakamura H (2012) Computer-aided antibody design. *Protein Engineering Design and Selection*.
17. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL (2004) Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J Med Chem* 47:1750-1759.
18. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, et al (2004) Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J Med Chem* 47:1739-1749.
19. Sircar A, Kim ET, Gray JJ (2009) RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res* 37:W474-W479.
20. Goodsell DS, Olson AJ (1990) Automated docking of substrates to proteins by simulated annealing. *Proteins* 8:195-202.

## COMBINED REFERENCES

- Adcock SA, & McCammon, J. A. (2006). Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chemical Reviews*, 106(5), 1589-1615.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403-410.
- Amoutzias, G., & Van de Peer, Y. (2010). Single-gene and whole-genome duplications and the evolution of protein-protein interaction networks. *Evolutionary genomics and systems biology* (pp. 413-429) John Wiley & Sons, Inc.  
doi:10.1002/9780470570418.ch19
- Andre, I., Strauss, C., Kaplan, D., Bradley, P., & Baker, D. (2008). *Emergence of symmetry in homooligomeric biological assemblies*
- Aytuna AS, Gursoy A, & Keskin, O. (2005). Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics (Oxford, England)*, 21(12), 2850-2855.
- Aytuna, A. S., Gursoy, A., & Keskin, O. (2005). Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics (Oxford, England)*, 21(12), 2850-2855.
- Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T., & Hogue, C. W. (2001). BIND--the biomolecular interaction network database. *Nucleic Acids Research*, 29(1), 242-245.
- Bahadur, R. P., & Zacharias, M. (2008). The interface of protein-protein complexes: Analysis of contacts and prediction of interactions. *Cellular and Molecular Life Sciences: CMLS*, 65(7-8), 1059-1072.
- Bahadur, R. P., Chakrabarti, P., Rodier, F., & Janin, J. (2003). Dissecting subunit interfaces in homodimeric proteins. *Proteins*, 53(3), 708-719.

- Bahadur, R. P., Chakrabarti, P., Rodier, F., & Janin, J. (2004). A dissection of specific and non-specific protein-protein interfaces. *Journal of Molecular Biology*, 336(4), 943-955.
- Ball, A. M. (2000). Gene ontology: Tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1), 25-29.
- Bartlett GJ, Porter CT, Borkakoti N, & Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *Journal of Molecular Biology*, 324(1), 105-121.
- Bastolla, U., Moya, A., Viguera, E., & van Ham, R. C. (2004). *Genomic determinants of protein folding thermodynamics in prokaryotic organisms*
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Research*, 28(1), 235-242.
- Berman, H. M., Battistuz, T., Bhat, T., Bluhm, W., Bourne, P., Burkhardt, K., et al. (2002). *The protein data bank*
- Bogan, A. A., & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*, 280(1), 1-9.
- Bordner, A. J., & Abagyan, R. (2005). Statistical analysis and prediction of protein-protein interfaces. *Proteins*, 60(3), 353-366.
- Bradshaw RT, Patel BH, Tate EW, Leatherbarrow RJ, & Gould, I. R. (2011). Comparing experimental and computational alanine scanning techniques for probing a prototypical protein-protein interaction. *Protein Engineering, Design & Selection: PEDS*, 24(1-2), 197-207.
- Brenke, R., Hall, D. R., Chuang, G., Comeau, S. R., Bohnuud, T., Beglov, D., et al. (2012). Application of asymmetric statistical potentials to antibody-protein docking. *Bioinformatics (Oxford, England)*, 28(20), 2608-2614. doi:10.1093/bioinformatics/bts493
- Brown, N. G., Chow, D., Ruprecht, K., & Palzkill, T. (2013). *Identification of the beta-lactamase inhibitor protein-II (BLIP-II) interface residues essential for binding affinity and specificity for class A beta-lactamases*
- Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J., & Huang, E. S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science: A Publication of the Protein Society*, 13(1), 190-202.

- Carugo, O., & Argos, P. (1997). *Protein-protein crystal-packing contacts*
- Chakrabarti, P., & Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins*, 47(3), 334-343.
- Chen, R., Li, L., & Weng, Z. (2003). ZDOCK: An initial-stage protein-docking algorithm. *Proteins*, 52(1), 80-87.
- Choi, Y. S., Yang, J., Choi, Y., Ryu, S. H., & Kim, S. (2009). Evolutionary conservation in multiple faces of protein interaction. *Proteins*, 77(1), 14-25.  
doi:10.1002/prot.22410
- De Las Rivas, J., & Fontanillo, C. (2010). Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Computational Biology*, 6(6), e1000807. doi:10.1371/journal.pcbi.1000807
- de Vries, S.,J., van Dijk, A.,D.J., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., et al. (2007). HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins*, 69(4), 726-733.
- del Sol, A., & O'Meara, P. (2005). Small-world network approach to identify key residues in protein-protein interaction. *Proteins*, 58(3), 672-682.
- DeLano, W. L. (2002). Unraveling hot spots in binding interfaces: Progress and challenges. *Current Opinion in Structural Biology*, 12(1), 14-20.
- Dey, S., Pal, A., Chakrabarti, P., & Janin, J. (2010a). The subunit interfaces of weakly associated homodimeric proteins. *Journal of Molecular Biology*, 398(1), 146-160.  
doi:10.1016/j.jmb.2010.02.020
- Dey, S., Pal, A., Chakrabarti, P., & Janin, J. (2010b). The subunit interfaces of weakly associated homodimeric proteins. *Journal of Molecular Biology*, 398(1), 146-160.  
doi:10.1016/j.jmb.2010.02.020
- Dominguez, C., Boelens, R., & Bonvin, A. M. J. J. (2003). HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7), 1731-1737. doi:10.1021/ja026939x
- Duarte, J. M., Srebniak, A., Schärer, M.,A., & Capitani, G. (2012). Protein interface classification by evolutionary analysis. *BMC Bioinformatics*, 13, 334-334.  
doi:10.1186/1471-2105-13-334

- Duran, A. M., & Meiler, J. (2013). Inverted topologies in membrane proteins: A mini-review. *Computational and Structural Biotechnology Journal*, 8, e201308004-e201308004.
- Dutta, S., & Berman, H. M. (2005). Large macromolecular complexes in the protein data bank: A status report. *Structure (London, England: 1993)*, 13(3), 381-388.
- Eisenstein, M., Shariv, I., Koren, G., Friesem, A. A., & Katchalski-Katzir, E. (1997). Modeling supra-molecular helices: Extension of the molecular surface recognition algorithm and application to the protein coat of the tobacco mosaic virus. *Journal of Molecular Biology*, 266(1), 135-143.
- Elcock AH, & McCammon, J. A. (2001). Identification of protein oligomerization states by analysis of interface conservation. *Proceedings of the National Academy of Sciences of the United States of America*, 98(6), 2990-2994.
- Erickson JA, Jalaie M, Robertson DH, Lewis RA, & Vieth, M. (2004). Lessons in molecular recognition: The effects of ligand and protein flexibility on molecular docking accuracy. *Journal of Medicinal Chemistry*, 47(1), 45-55.
- Ezkurdia, I., Bartoli, L., Fariselli, P., Casadio, R., Valencia, A., & Tress, M. L. (2009). Progress and challenges in predicting protein-protein interaction sites. *Briefings in Bioinformatics*, 10(3), 233-246. doi:10.1093/bib/bbp021
- Fernández-Recio, J., Totrov, M., & Abagyan, R. (2003). ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins*, 52(1), 113-117.
- Fink, F., Ederer, S., & Gronwald, W. (2009). Protein-protein interaction analysis by docking. *Algorithms*, 2, 429.
- Fischer, T. B., Arunachalam, K. V., Bailey, D., Mangual, V., Bakhru, S., Russo, R., et al. (2003). The binding interface database (BID): A compilation of amino acid hot spots in protein interfaces. *Bioinformatics (Oxford, England)*, 19(11), 1453-1454.
- Fleishman, S. J., Whitehead, T. A., Strauch, E., Corn, J. E., Qin, S., Zhou, H., et al. (2011). Community-wide assessment of protein-interface modeling suggests improvements to design methodology. *Journal of Molecular Biology*, 414(2), 289-302. doi:10.1016/j.jmb.2011.09.031
- Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7), 1739-1749. doi:10.1021/jm0306430

- Gabb, H. A., Jackson, R. M., & Sternberg, M. J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of Molecular Biology*, 272(1), 106-120.
- Garzon, J. I., López-Blanco, J. R., Pons, C., Kovacs, J., Abagyan, R., Fernandez-Recio, J., et al. (2009). FRODOCK: A new approach for fast rotational protein-protein docking. *Bioinformatics (Oxford, England)*, 25(19), 2544-2551.  
doi:10.1093/bioinformatics/btp447
- Gavin, A., Bösch, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868), 141-147.
- Gohlke, H., Hendlich, M., & Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. *Journal of Molecular Biology*, 295(2), 337-356.
- González-Ruiz, D., & Gohlke, H. (2006). Targeting protein-protein interactions with small molecules: Challenges and perspectives for computational binding epitope detection and ligand finding. *Current Medicinal Chemistry*, 13(22), 2607-2625.
- Goodsell, D. S., & Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins*, 8(3), 195-202.
- Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., et al. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*, 331(1), 281-299.
- Greenbaum, J. A., Andersen, P., Blythe, M., Bui, H., Cachau, R., Crowe, J., et al. (2007). *Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools*
- Grishin NV, & Phillips, M. A. (1994). The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Science: A Publication of the Protein Society*, 3(12), 2455-2458.
- Grishin, N. V., Phillips, M. A., & Goldsmith, E. J. (1995). Modeling of the spatial structure of eukaryotic ornithine decarboxylases. *Protein Science: A Publication of the Protein Society*, 4(7), 1291-1304.

- Guharoy, M., & Chakrabarti, P. (2005). Conservation and relative importance of residues across protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15447-15452.
- Guharoy, M., & Chakrabarti, P. (2010). Conserved residue clusters at protein-protein interfaces and their use in binding site identification. *BMC Bioinformatics*, 11, 286-286. doi:10.1186/1471-2105-11-286
- Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., et al. (2004). Glide: A new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *Journal of Medicinal Chemistry*, 47(7), 1750-1759. doi:10.1021/jm030644s
- Halperin, I., Ma, B., Wolfson, H., & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics*, 47(4), 409-443. doi:10.1002/prot.10115
- Harris, R., Olson, A. J., & Goodsell, D. S. (2008). Automated prediction of ligand-binding sites in proteins. *Proteins*, 70(4), 1506-1517.
- Hartigan, J. A. (1973). Clustering. *Annual Review of Biophysics and Bioengineering*, 2, 81-101. doi:10.1146/annurev.bb.02.060173.000501
- Havranek, J. J. (2010). Specificity in computational protein design. *The Journal of Biological Chemistry*, 285(41), 31095-31099. doi:10.1074/jbc.R110.157685
- Hegyi, H., & Gerstein, M. (1999). *The relationship between protein structure and function: A comprehensive survey with application to the yeast genome*
- Henrick, K. E. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, 372(3), 774-797.
- Henrick, P. H. (2000). Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, 41(1), 47-57.
- Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S., et al. (2002). Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868), 180-183.
- Honig, B., & Yang, A. S. (1995). Free energy balance in protein folding. *Advances in Protein Chemistry*, 46, 27-58.

Hsin, K., Ghosh, S., & Kitano, H. (2013). Combining machine learning systems and multiple docking simulation packages to improve docking prediction reliability for network pharmacology. *Plos One*, 8(12), e83922-e83922. doi:10.1371/journal.pone.0083922

[Http://Mammoth.bcm.tmc.edu](http://Mammoth.bcm.tmc.edu).

[Http://Robetta.bakerlab.org](http://Robetta.bakerlab.org).

Hubbard, S. J., & Argos, P. (1994). Cavities and packing at protein interfaces. *Protein Science: A Publication of the Protein Society*, 3(12), 2194-2206.

Hubbard, T. J., Ailey, B., Brenner, S. E., Murzin, A. G., & Chothia, C. (1999). SCOP: A structural classification of proteins database. *Nucleic Acids Research*, 27(1), 254-256.

Ishchenko, A. V., & Shakhnovich, E. I. (2002). SMOG2001: An improved knowledge-based scoring function for protein-ligand interactions. *Journal of Medicinal Chemistry*, 45(13), 2770-2780.

Jackson, R. M., Gabb, H. A., & Sternberg, M. J. (1998). Rapid refinement of protein interfaces incorporating solvation: Application to the docking problem. *Journal of Molecular Biology*, 276(1), 265-285.

Janin, J., & Rodier, F. (1995). *Protein-protein interaction at crystal contacts*

Janin, J. (2005). Assessing predictions of protein-protein interaction: The CAPRI experiment. *Protein Science: A Publication of the Protein Society*, 14(2), 278-283.

Janin, J. (2010). Protein-protein docking tested in blind predictions: The CAPRI experiment. *Molecular Biosystems*, 6(12), 2351-2362. doi:10.1039/c005060c

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J. E., Vajda, S., et al. (2003a). CAPRI: A critical assessment of PRedicted interactions. *Proteins*, 52(1), 2-9.

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J. E., Vajda, S., et al. (2003b). CAPRI: A critical assessment of PRedicted interactions. *Proteins*, 52(1), 2-9.

Jones, S., & Thornton, J. M. (1995). Protein-protein interactions: A review of protein dimer structures. *Progress in Biophysics and Molecular Biology*, 63(1), 31-65.

- Jones, S., & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 93(1), 13-20.
- Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., & Vakser, I. A. (1992). Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences of the United States of America*, 89(6), 2195-2199.
- Keskin, O., Gursoy, A., Ma, B., & Nussinov, R. (2008). Principles of protein-protein interactions: What are the preferred ways for proteins to interact? *Chemical Reviews*, 108(4), 1225-1244. doi:10.1021/cr040409x
- Keskin, O., Ma, B., & Nussinov, R. (2005). Hot regions in protein--protein interactions: The organization and contribution of structurally conserved hot spot residues. *Journal of Molecular Biology*, 345(5), 1281-1294.
- Kim, D. E., Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the robetta server. *Nucleic Acids Research*, 32, W526-W531.
- Kim, K. T. (2004). Computational alanine scanning of protein-protein interfaces. *Science's STKE: Signal Transduction Knowledge Environment*, 2004(219), pl2-pl2.
- Kortemme, T., & Baker, D. (2002). A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22), 14116-14121.
- Kozakov, D., Brenke, R., Comeau, S., & Vajda, S. (2006). *PIPER: An FFT-based protein docking program with pairwise potentials*
- Krissinel, E. (2010). Crystal contacts as nature's docking solutions. *Journal of Computational Chemistry*, 31(1), 133-143.
- Krissinel, E., & Henrick, K. (2005). Detection of protein assemblies in crystals. In M. R. Berthold, R. Glen, K. Diederichs, O. Kohlbacher & I. Fischer (Eds.), (pp. 163-174) Springer Berlin Heidelberg. doi:10.1007/11560500\_15
- Krüger, D., M., Ignacio Garzón, J., Chacón, P., & Gohlke, H. (2014). DrugScorePPI knowledge-based potentials used as scoring and objective function in protein-protein docking. *Plos One*, 9(2), e89466-e89466. doi:10.1371/journal.pone.0089466

- Kuroda, D., Shirai, H., Jacobson, M. P., & Nakamura, H. (2012). Computer-aided antibody design. *Protein Engineering Design and Selection*, doi:10.1093/protein/gzs024
- Larsen, T. A., Olson, A. J., & Goodsell, D. S. (1998). Morphology of protein-protein interfaces. *Structure (London, England: 1993)*, 6(4), 421-427.
- Laskowski, R. A. (1995). SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics*, 13(5), 323.
- Levy, E. D. (2007). PiQSi: Protein quaternary structure investigation. *Structure (London, England: 1993)*, 15(11), 1364-1367.
- Li, B., & Kihara, D. (2012). Protein docking prediction using predicted protein-protein interface. *BMC Bioinformatics*, 13, 7-7. doi:10.1186/1471-2105-13-7
- Li, L., Chen, R., & Weng, Z. (2003). RDOCK: Refinement of rigid-body protein docking predictions. *Proteins: Structure, Function, and Bioinformatics*, 53(3), 693-707. doi:10.1002/prot.10460
- Li, L., Guo, D., Huang, Y., Liu, S., & Xiao, Y. (2011). ASPDock: Protein-protein docking algorithm using atomic solvation parameters model. *BMC Bioinformatics*, 12, 36-36. doi:10.1186/1471-2105-12-36
- Lichtarge, O., Bourne, H. R., & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology*, 257(2), 342-358.
- Liu, S., Li, Q., & Lai, L. (2006). A combinatorial score to distinguish biological and nonbiological protein-protein interfaces. *Proteins*, 64(1), 68-78.
- Lo Conte, L., Chothia, C., & Janin, J. (1999a). The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*, 285(5), 2177-2198.
- Lo Conte, L., Chothia, C., & Janin, J. (1999b). The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*, 285(5), 2177-2198.
- Lockless, S. W., & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science (New York, N.Y.)*, 286(5438), 295-299.

- Lopes, A., Sacquin-Mora, S., Dimitrova, V., Laine, E., Ponty, Y., & Carbone, A. (2013). Protein-protein interactions in a crowded environment: An analysis via cross-docking simulations and evolutionary information. *Plos Computational Biology*, 9(12), e1003369-e1003369. doi:10.1371/journal.pcbi.1003369
- Ma, B., Elkayam, T., Wolfson, H., & Nussinov, R. (2003a). Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10), 5772-5777.
- Ma, B., Elkayam, T., Wolfson, H., & Nussinov, R. (2003b). Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10), 5772-5777.
- Madaoui, H., & Guerois, R. (2008). Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proceedings of the National Academy of Sciences*, 105(22), 7708-7713. doi:10.1073/pnas.0707032105
- Malod-Dognin, N., Bansal, A., & Cazals, F. (2012). Characterizing the morphology of protein binding patches. *Proteins*, 80(12), 2652-2665. doi:10.1002/prot.24144
- Massova, I., & Kollman, P. A. (1999). Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *Journal of the American Chemical Society*, 121(36), 8133-8143. doi:10.1021/ja990935j
- Meenan, N. A. G., Sharma, A., Fleishman, S. J., Macdonald, C. J., Morel, B., Boetzel, R., et al. (2010). The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 107(22), 10080-10085. doi:10.1073/pnas.0910756107
- Mihalek, I., Reš, I., & Lichtarge, O. (2004). A family of Evolution–Entropy hybrid methods for ranking protein residues by importance. *Journal of Molecular Biology*, 336(5), 1265. doi:10.1016/j.jmb.2003.12.078
- Mihalek, I., Res, I., & Lichtarge, O. (2006). Evolutionary trace report\_maker: A new type of service for comparative analysis of proteins. *Bioinformatics (Oxford, England)*, 22(13), 1656-1657.

- Mintseris, J., & Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31), 10930-10935.
- Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J., et al. (2005). Protein-protein docking benchmark 2.0: An update. *Proteins*, 60(2), 214-216.
- Mishra, S. (2012). Computational prediction of protein-protein complexes. *BMC Research Notes*, 5, 495-495. doi:10.1186/1756-0500-5-495
- Monsellier, E., & Chiti, F. (2007). *Prevention of amyloid-like aggregation as a driving force of protein evolution*
- Moreira, I. S., Fernandes, P. A., & Ramos, M. J. (2010). Protein-protein docking dealing with the unknown. *Journal of Computational Chemistry*, 31(2), 317-342. doi:10.1002/jcc.21276
- Moretti, R., Fleishman, S. J., Agius, R., Torchala, M., Bates, P. A., Kastiris, P. L., et al. (2013). Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. *Proteins*,
- Morgan, D. H., Kristensen, D. M., Mittelman, D., & Lichtarge, O. (2006). ET viewer: An application for predicting and visualizing functional sites in protein structures. *Bioinformatics (Oxford, England)*, 22(16), 2049-2050.
- Morrison, K. L., & Weiss, G. A. (2001). Combinatorial alanine-scanning. *Current Opinion in Chemical Biology*, 5(3), 302-307.
- Muegge, I., & Martin, Y. C. (1999). A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *Journal of Medicinal Chemistry*, 42(5), 791-804.
- Neuvirth, H., Raz, R., & Schreiber, G. (2004). ProMate: A structure based prediction program to identify the location of protein-protein binding sites. *Journal of Molecular Biology*, 338(1), 181-199.
- Nooren IM, & Thornton, J. M. (2003). Diversity of protein-protein interactions. *The EMBO Journal*, 22(14), 3486-3492.
- Nyfeler, B., Michnick, S. W., & Hauri, H. (2005). Capturing protein interactions in the secretory pathway of living cells. *Proceedings of the National Academy of Sciences of the United States of America*, 102(18), 6350-6355.

- Ofran, Y., & Rost, B. (2003). Analysing six types of protein-protein interfaces. *Journal of Molecular Biology*, 325(2), 377-387.
- Ofran, Y., & Rost, B. (2007). Protein-protein interaction hotspots carved into sequences. *Plos Computational Biology*, 3(7), e119-e119.
- Ohta, T. (1976). *Role of very slightly deleterious mutations in molecular evolution and polymorphism*
- Olson, A. J., & Goodsell, D. S. (1998). *Automated docking and the search for HIV protease inhibitors*
- Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure (London, England: 1993)*, 5(8), 1093-1108.
- Ouzounis, C. F., Perez-Irratxeta, C. F., Sander, C. F., & Valencia, A. (1998). *Are binding residues conserved?*
- Palma, P. N., Krippahl, L., Wampler, J., & Moura, J. (2000). *BiGGER: A new (soft) docking algorithm for predicting protein interactions*
- Panchenko, A. R., Kondrashov, F., & Bryant, S. (2004). Prediction of functional sites by analysis of sequence and structure conservation. *Protein Science*, 13(4), 884-892. doi:10.1110/ps.03465504
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2, 11, 559-572.
- Perola, E. (2006). *Minimizing false positives in kinase virtual screens*
- Phizicky EM, & Fields, S. (1995). Protein-protein interactions: Methods for detection and analysis. *Microbiological Reviews*, 59(1), 94-123.
- Ritchie DW, & Kemp, G. J. (2000). Protein docking using spherical polar fourier correlations. *Proteins*, 39(2), 178-194.
- Ritchie, D. W., & Kemp, G. J. (2000). Protein docking using spherical polar fourier correlations. *Proteins*, 39(2), 178-194.
- SAS Institute Inc., Cary, NC. (1989-2012). *JMP*

- Scharer, M. A., Grutter, M., & Capitani, G. (2010). *CRK: An evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts*
- Schatz, D. G., & Swanson, P. C. (2011). V(D)J recombination: Mechanisms of initiation. *Annual Review of Genetics*, 45, 167-202. doi:10.1146/annurev-genet-110410-132552
- Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., & Wolfson, H. J. (2005). PatchDock and SymmDock: Servers for rigid and symmetric docking. *Nucleic Acids Research*, 33, W363-W367.
- Schrödinger, L. *The PyMOL molecular graphics system*
- Schwartz, G. W., & Hershberg, U. (2013). Germline amino acid diversity in B cell receptors is a good predictor of somatic selection pressures. *Frontiers in Immunology*, 4, 357-357. doi:10.3389/fimmu.2013.00357
- Shoichet BK, McGovern SL, Wei B, & Irwin, J. J. (2002). Lead discovery using molecular docking. *Current Opinion in Chemical Biology*, 6(4), 439-446.
- Sircar, A., Kim, E. T., & Gray, J. J. (2009). RosettaAntibody: Antibody variable region homology modeling server. *Nucleic Acids Research*, 37, W474-W479. doi:10.1093/nar/gkp387
- Smith GR, & Sternberg, M. J. (2002). Prediction of protein-protein interactions by docking methods. *Current Opinion in Structural Biology*, 12(1), 28-35.
- Sonavane, S., & Chakrabarti, P. (2008). Cavities and atomic packing in protein structures and interfaces. *Plos Computational Biology*, 4(9), e1000188-e1000188. doi:10.1371/journal.pcbi.1000188
- Søndergaard, C., R., Garrett, A. E., Carstensen, T., Pollastri, G., & Nielsen, J. E. (2009). Structural artifacts in protein-ligand X-ray structures: Implications for the development of docking scoring functions. *Journal of Medicinal Chemistry*, 52(18), 5673-5684. doi:10.1021/jm8016464
- Sudarshan, S., & Beck, B. (In Revision). <br />Functional classification of protein interactions using interface spatial distribution of evolutionary criteria.
- Sudarshan, S., Kodathala, S. B., Mahadik, A. C., Mehta, I., & Beck, B. W. (2014). Protein-protein interface detection using the energy centrality relationship (ECR) characteristic of proteins. *Plos One*, 9(5), e97115-e97115. doi:10.1371/journal.pone.0097115

- Swapna, L. S., Bhaskara, R. M., Sharma, J., & Srinivasan, N. (2012). Roles of residues in the interface of transient protein-protein complexes before complexation. *Scientific Reports*, 2, 334-334. doi:10.1038/srep00334
- Szilágyi, A., Grimm, V., Arakaki, A. K., & Skolnick, J. (2005). Prediction of physical protein-protein interactions. *Physical Biology*, 2(2), S1-S16.
- Talavera, D., Robertson, D., & Lovell, S. (2011). *Characterization of protein-protein interaction interfaces from a single species*
- Ten Eyck, L. F., Mandell, J., Roberts, V. A., & Pique, M. E. (1995). Surveying molecular interactions with DOT. Paper presented at the *Supercomputing, 1995. Proceedings of the IEEE/ACM SC95 Conference*, pp. 22-22. doi:10.1109/SUPERC.1995.242670
- Thorn, K. S., & Bogan, A. A. (2001). ASEdb: A database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17(3), 284-285. doi:10.1093/bioinformatics/17.3.284
- Tsai, C. J., Lin, S. L., Wolfson, H. J., & Nussinov, R. (1996). A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *Journal of Molecular Biology*, 260(4), 604-620.
- Tuncbag, N., Kar, G., Keskin, O., Gursoy, A., & Nussinov, R. (2009). A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Briefings in Bioinformatics*, 10(3), 217-232. doi:10.1093/bib/bbp001
- Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., et al. (2000). A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae*. *Nature*, 403(6770), 623-627.
- Vakser, I. A. (1997). Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins, Suppl 1*, 226-230.
- Valdar WS, & Thornton, J. M. (2001a). Conservation helps to identify biologically relevant crystal contacts. *Journal of Molecular Biology*, 313(2), 399-416.
- Valdar WS, & Thornton, J. M. (2001b). Protein-protein interfaces: Analysis of amino acid conservation in homodimers. *Proteins*, 42(1), 108-124.
- Valdar, W. S., & Thornton, J. M. (2001). Protein-protein interfaces: Analysis of amino acid conservation in homodimers. *Proteins*, 42(1), 108-124.

- Valdar, W. S. (2002). Scoring residue conservation. *Proteins*, 48(2), 227-241.
- Valencia, A., & Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology*, 12(3), 368-373.
- Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., & Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins*, 52(4), 609-623.
- Verdonk, M. L., Giangreco, I., Hall, R. J., Korb, O., Mortenson, P. N., & Murray, C. W. (2011). Docking performance of fragments and druglike compounds. *Journal of Medicinal Chemistry*, 54(15), 5422-5431. doi:10.1021/jm200558u
- Warren, G. L., Andrews, C. W., Capelli, A., Clarke, B., LaLonde, J., Lambert, M. H., et al. (2006). A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry*, 49(20), 5912-5931.
- Weiss, M. S., Metzner, H. J., & Hilgenfeld, R. (1998). Two non-proline cis peptide bonds may be important for factor XIII function. *FEBS Letters*, 423(3), 291-296.
- Wells, C. T. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science (New York, N.Y.)*, 267(5196), 383-386.
- Williams, T., & Kelley, C. (2011).
- Wodak, S. J., & Janin, J. (1978). Computer analysis of protein-protein interaction. *Journal of Molecular Biology*, 124(2), 323-342.
- Wodak, S. J., & Janin, J. (2002). Structural basis of macromolecular recognition. *Advances in Protein Chemistry*, 61, 9-73.
- Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., & Eisenberg, D. (2002). DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1), 303-305.
- Young, K. H. (1998). Yeast two-hybrid: So many interactions, (in) so little time.. *Biology of Reproduction*, 58(2), 302-311.
- Young, L., Jernigan, R. L., & Covell, D. G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Protein Science: A Publication of the Protein Society*, 3(5), 717-729.

- Zacharias, M. (2005). ATTRACT: Protein-protein docking in CAPRI using a reduced protein model. *Proteins*, 60(2), 252-256.
- Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., & Cesareni, G. (2002). MINT: A molecular INTERaction database. *FEBS Letters*, 513(1), 135-140.
- Zhu, H., Domingues, F. S., Sommer, I., & Lengauer, T. (2006). NOXclass: Prediction of protein-protein interaction types. *BMC Bioinformatics*, 7, 27-27.

APPENDIX  
USING THE ENERGY CENTRALITY RELATIONSHIP IN A CAPRI-TYPE  
ANALYSIS OF UNBOUND TARGETS

## **Abstract**

CAPRI (Critical Assessment of Prediction of Interactions) is a community-wide experiment that aims to evaluate methods that model multi-component molecular interactions, including proteins and nucleic acids<sup>1</sup>. While a number of methods are being used and developed to predict these interactions, their success remains limited especially when predicting interactions between unbound components. The success of ECR and its ability to improve bound docking runs led us to assess ECR's success on unbound structures (structures that could undergo conformational change upon binding). Here we use unbound docking targets from previous CAPRI analyses to evaluate the efficiency of ECR in predicting protein interactions complexes. HEX was unsuccessful in generating any native-like docking poses for the targets selected. As a result, ECR was not able to identify any native-like poses either.

## **Introduction**

The structure of a Protein-Protein complex can provide details about its biological role<sup>2</sup>. Structural information obtained from techniques such as NMR and X-ray crystallography is deposited in the Protein Data Bank (PDB)<sup>3</sup>. However, PDB has a relative dearth of large complexes<sup>4</sup> and the structures available may not always represent the biological assembly of the complex<sup>5</sup>. In the absence of accurate three-dimensional protein interaction data, computational methods can be used to generate structures.

Protein-protein docking is one such computational technique that addresses the

problem of deriving three-dimensional structures of complexes, starting from the atomic positions of the individual components<sup>6,7</sup>. Docking can provide insights into molecular functions of proteins based on properties such as the affinity and specificity of interaction<sup>8</sup>. Docking methods developed in the recent past are successful in generating near-native conformations. Unfortunately, they also generate a large number of non-near native structures for every successful prediction (false positives)<sup>9</sup>. As docking algorithms have improved to overcome these and other drawbacks, tools to blindly assess and compare their predictive capacity have become necessary<sup>9</sup>.

The Critical Assessment of Protein Interactions (CAPRI) is a community wide experiment designed to assess the reliability of docking in structure prediction. Each prediction round has 1-7 experimentally determined quaternary structures. The coordinates of these structures are not provided to participants and blind predictions are therefore submitted by participants and assessed by judges<sup>9</sup>. In Chapter 4, we showed docking analysis on bound conformations of proteins (conformations already in an oligomeric form) to validate ECR's ability to identify native-like structures. Bound docking involves dissociating the components of a known complex and reassembling the complex by docking. Unbound docking is similar, but involves assembling the known complex from known structures of the isolated, "unbound", component proteins. It is not generally likely that the isolated proteins will have exactly the same structure as their final complexed conformations. Fully unbound CAPRI targets are generally desirable, as the crystal structure of bound targets may occur in a highly specific conformation that

may not sufficiently represent the unbound target. However, as fully unbound protein targets are not abundant, mixed “bound/unbound” targets are also sometimes assessed <sup>9</sup>.

As the ability of our ECR method to identify native-like structures in interactions involving an unbound component has not been determined, we identified potential target structures from previous CAPRI experiments. We actively avoided protein-nucleotide interactions in order to concentrate on protein:protein interactions (PPI). Interactions requiring homology modeling or molecular dynamics simulation were also avoided, as such studies are much more computationally expensive. Given these restrictions, we narrowed the list of 53 CAPRI targets to 3 potential test cases. These cases involved a regulator (Target 30), and two enzyme-inhibitors (Target 40 and 41). Since ECR previously had trouble identifying AbAg (Antibody-Antigen) and Enzyme-Inhibitor interactions, targets involving such interactions have the potential to be challenging cases. Additionally, our Docking/ECR studies (Chapter 4) were performed on bound conformations of proteins; interfaces involving movement or conformational rearrangement (as in unbound components) had the potential to prove to be even more problematic. Regardless, we chose to proceed, as even if our predictions were inaccurate, they would provide information that would enable us to build a better methodology.

## Results

### ***Target 30: Rnd1-GTP in complex with plexin B1.***

The unbound structures provided by CAPRI were PDBid *2cls*, chain A (Rho-related GTP binding protein) and, PDBid *2r2o*, chain B (plexin-B1 protein). The complex to be predicted was PDBid *2rex* (identified after the Target 20 round had completed). All structures are shown in Figure A1a, b, and c.

Poses generated by HEX were analyzed by ECR (Figure A1.2). Only solution 913 from HEX was identified in the FLIP cluster (Figure A1.1d, A1.2). This structure had a root mean square deviation (RMSD) of 48.31 Å to the known structure (where native-like is defined to be  $\leq 5$  Å). In addition, no solution matched any symmetry-transformed conformations of *2rex*. Neither HEX nor HEX+ECR were able to identify a native-like pose.

### ***Target 40: Double headed inhibitor API-A bound to two trypsin molecules.***

The structures provided were the bound form of a protease inhibitor, PDBid *3e8l*, chain C in the bound form and unbound bovine trypsin PDBid *1bty*. The structure to predict was the C:B interaction of *3e8l*. The question being posed was could the unbound structure of trypsin be used to identify the bound form of trypsin with an inhibitor. All structures are shown in Figure A1.3a, b, and c.

When the poses generated by HEX were analyzed by ECR (Figure A1.4), we were able to identify many poses in the FLIP region from a second clustering analysis. We identified solution 530 from HEX as the ECR solution (Figures A1.3d, A1.4). However, both the HEX best structure (solution 72, RMSD 7.91 Å) and ECR best structure (solution 530, RMSD 54.69 Å, Figure A1.4) failed to identify a native-like structure.

***Target 41: Colicin E9 dnase domain in complex with IM2 protein.***

The structures provided were unbound colicin, PDBid *lfsj*, chain B and the unbound immunity protein PDBid *Ino8*, chain A. The target structure was PDBid *2wpt*. All structures are shown in Figures A1.5a, b and c.

The poses generated by HEX were analyzed by ECR (Figure A1.6) and only solution 823 from HEX was identified in the FLIP cluster (Figures A1.5d, A1.6). This structure had an RMSD of 30.56 Å and also did not match any symmetry-transformed conformations of *2wpt*. Once again HEX, and as a result ECR, failed to identify a native-like structure.

**Discussion**

The docking poses of all three targets, particularly the non-native poses, exhibited similar PCA distributions mostly occupying the negative PC1 region of quadrant 2. This pattern of distribution of non-native docking poses was also seen in our analysis of bound-bound interactions in Chapter 4. We take this to suggest that properties specific to

non-native docking poses, but neither FunC or FLIP structures, continue to be identified by ECR, even in the presence of unbound source structures. As non-native docking poses cluster nearer to the FunC cluster, we suggest this indicates that they are more similar in physical properties as well. It is possible that docking software like HEX emphasizes affinity over specificity and generates structures that may be stable complexes but not represent specific interactions necessary for function.

It is unclear how effective using ECR as a post-filter on docking poses with no native-like conformations would be, as we can not currently define what the most FLIP-like but non-near native structure means, if it has any significance at all. HEX failed to generate any native-like poses and, as expected, ECR therefore also couldn't identify any native-like poses. This lack of success could be because of the following reasons:

1. *Limited movement of interfacial residues.*

HEX and ECR do well in bound-bound predictions (Chapter 4). However, it is possible this is due to limited movement of the interfacial residues. It remains to be seen whether ECR is sensitive to induced-fit conformational dynamics.

Rearrangement of residues at the interface during PPI formation and the limited amount of conformational relaxation of residues by HEX could be introducing (and failing to relieve) shape/Van der Waals and electrostatic clashes. This in turn could lead to less favorable overall binding energies and result in rejection of otherwise native-like conformations that could have been identified using ECR. A potential

avenue for correcting this may be to identify if native-like, but poorly scoring ECR targets, have additional features that can be exploited for identification purposes.

## 2. *Lack of a well-defined FunC definition.*

ECR identifies the pose producing the most positive PC1 as its FLIP solution. This definition of ECR's FLIP solution is not particularly adept at identifying enzyme-inhibitor complexes. In earlier work, we have observed such interfaces to have FunC-like properties, which increases the likelihood of such poses being mistaken for false negatives. Future work needs to more rigorously/mathematically define FunC characteristics, as well as those of the more FunC-like FLIPs.

## **Conclusion**

In summary, it remains to be seen whether ECR (trained on bound protein structures) is an effective post-filter for unbound docking. ECR's effectiveness as a post-filter is limited by the results of the docking program used. It may be useful to identify docking programs dedicated to particularly challenging sub-categories such as AbAg and enzyme-inhibitors that HEX (Chapter 4) and ECR had limited success in predicting (Chapter 2).

## **Methods**

### ***Identifying targets***

Three potential targets were identified from previous CAPRI experiments. Protein-nucleotide interactions were discounted to concentrate on PPIs and interactions

requiring homology modeling or molecular dynamics simulation were excluded to avoid expensive computation. The three docking targets identified were Target 30 (a regulatory protein interaction), Target 40 (an Enzyme-Inhibitor complex), and Target 41 (also an Enzyme-Inhibitor complex).

### ***Molecular docking***

The protein docking program HEX was used to generate poses for protein structures in FLIPlite. Shape and electrostatic contributions to the docking correlation were enabled. The program was set to generate 1000 poses at Euler rotational increments of 7 degrees and a twist of 2 degrees. Chain 1 of the dimer was always identified as the receptor, which was fixed, and chain 2 was identified as the ligand, which sampled conformational space around the receptor. For each docking pose generated, an all atom RMSD to the PDB structure was calculated.

### ***Computational alanine scanning (CAS)***

The CAS method of Kortemme and Baker<sup>10, 11</sup> was used to process all the interfaces in the FLIPdb. In brief, this method evaluates enthalpy and free energy of solvation terms over conformations arising from a rotamer library for both the existing and alanine substituted residues in a PPI (native Gly and Pro excluded). These terms are used to determine a pseudo-free energy change upon substitution ( $\Delta\Delta G$ )<sup>11</sup>.

### ***Interfacial geometry***

Interfacial residues were defined using the same interface definition as in the CAS method of Kortemme and Baker<sup>11</sup>. The geometric distribution of residues in each PPI were determined by calculating the displacement ( $\Delta r$ ) of the C $\alpha$  position from the mean of the C $\alpha$  positions (termed the Center of Interface, CoI) using software written by the authors. A linear regression of the  $\Delta\Delta G$  and  $\Delta r$  data to a first-order polynomial ( $\Delta\Delta G = \text{slope} * \Delta r + \text{intercept}$ ) was calculated for each interface using software written by the authors as well as GNUPLOT<sup>12</sup>.

### ***Energetic and geometric features***

Calculations used in this work followed our previous protocol<sup>13</sup>, identifying 7 features for each interface: the slope (slope\_ $\Delta\Delta G$ ), intercept (intcpt\_ $\Delta\Delta G$ ), net sum of all  $\Delta\Delta G$  changes (sum\_ $\Delta\Delta G$ ), mean  $\Delta\Delta G$  for all interface residues (avg\_ $\Delta\Delta G$ ), total number of residues in the interface (#total), number of residues with  $\Delta\Delta G$  larger than +1 kcal/mol (#hot), and the ratio of “hot” to total (frac\_hot).

### ***Principal component analysis (PCA)***

Principal Component Analysis of the variation of CAS energetic and geometric feature data for all PPI was undertaken using JMP<sup>14</sup>. PCA determines a set of linearly-uncoupled eigenvectors from normalized correlations between variables that progressively describe the largest sources of variance in a data set<sup>15</sup>.

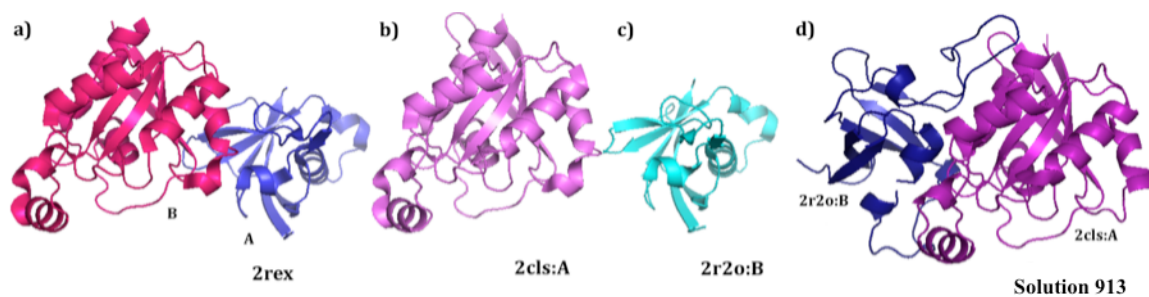
### ***K-means clustering***

K-means clustering<sup>16</sup> is a data analysis method that clusters observations into a specific number of clusters by attempting to find the point(s) that have the lowest mean variation from the other input data. When combined with PCA, the combination of features that allows input data to be clustered can be identified. In this work, we projected our energy versus distance data for the poses of the 3 proteins through a polynomial obtained from our earlier work on the training set<sup>13</sup>. ECR's pick was the structure with the most positive PC1 in the FLIP region of our clustering analysis.

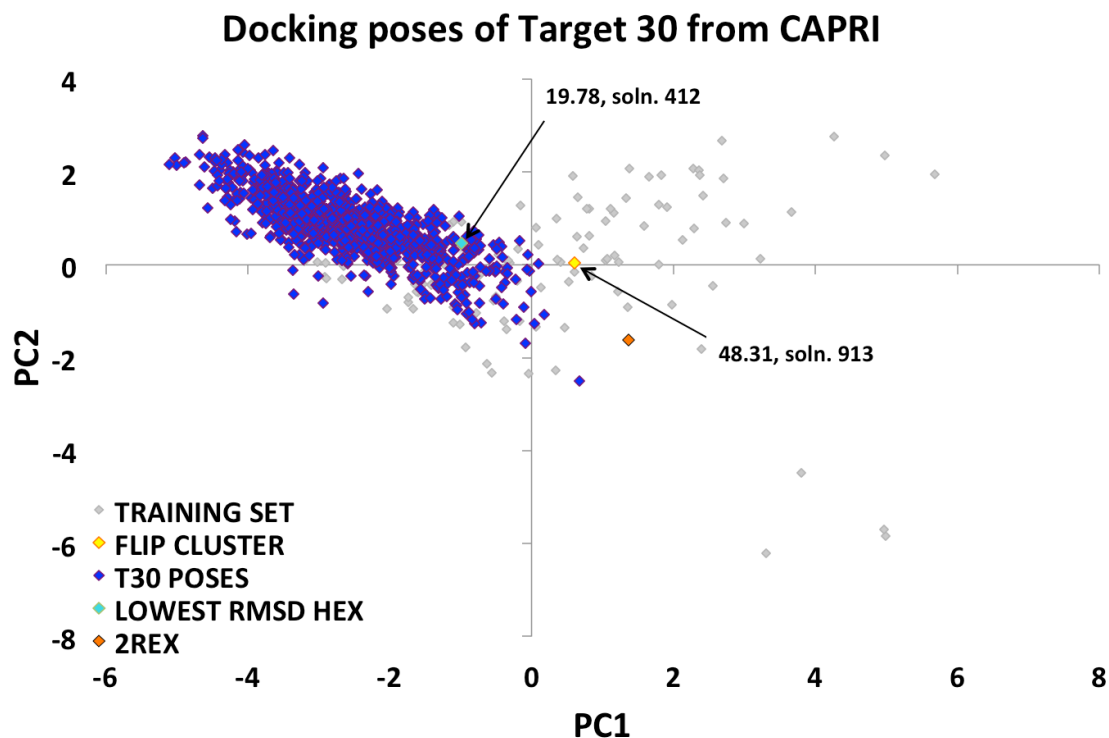
## References

1. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJE, Vajda S, Vakser I, Wodak SJ (2003) CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* 52:2-9.
2. Hegyi H, Gerstein M (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *Journal of Molecular Biology* JID - 2985088R.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235-242.
4. Berman HM, Battistuz T, Bhat T, Bluhm W, Bourne P, Burkhardt K, Feng Z, Gilliland G, Iype L, Jain S, et al (2002) The Protein Data Bank. *Acta Crystallographica. Section D, Biological Crystallography* JID - 9305878.
5. Dutta S, Berman HM (2005) Large macromolecular complexes in the Protein Data Bank: a status report. *Structure* 13:381-388.
6. Wodak SJ, Janin J (1978) Computer analysis of protein-protein interaction. *J Mol Biol* 124:323-342.
7. Wodak SJ, Janin J (2002) Structural basis of macromolecular recognition. *Adv Protein Chem* 61:9-73.
8. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* 33:W363-W367.
9. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJE, Vajda S, Vakser I, Wodak SJ (2003) CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* 52:2-9.
10. Kortemme T, Baker D (2002) A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A* 99:14116-14121.
11. Kim KT (2004) Computational alanine scanning of protein-protein interfaces. *Sci STKE* 2004:pl2-pl2.
12. Williams T, Kelley C (2011) .
13. Sudarshan S, Kodathala SB, Mahadik AC, Mehta I, Beck BW (2014) Protein-protein interface detection using the energy centrality relationship (ECR) characteristic of proteins. *PLoS One* 9:e97115-e97115.

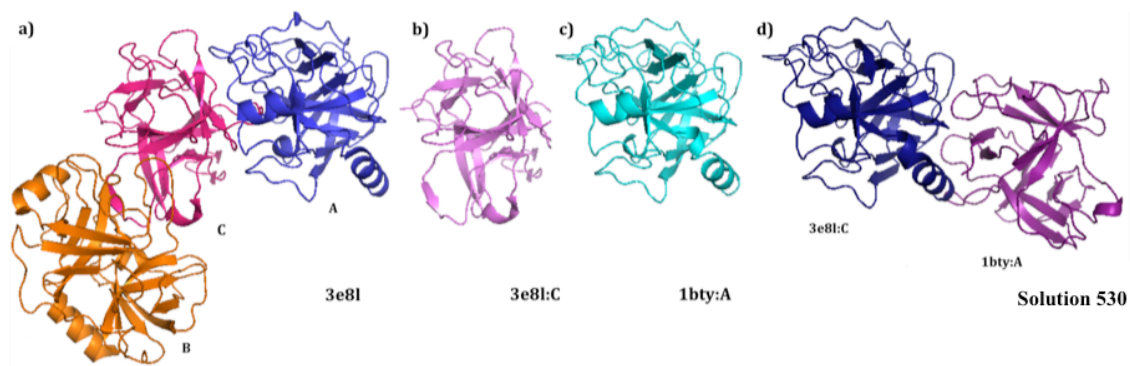
14. SAS Institute Inc., Cary, NC (1989-2012) JMP. 10.
15. Pearson K (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine 2 11:559-572.
16. Hartigan JA (1973) Clustering. Annu Rev Biophys Bioeng 2:81-101.



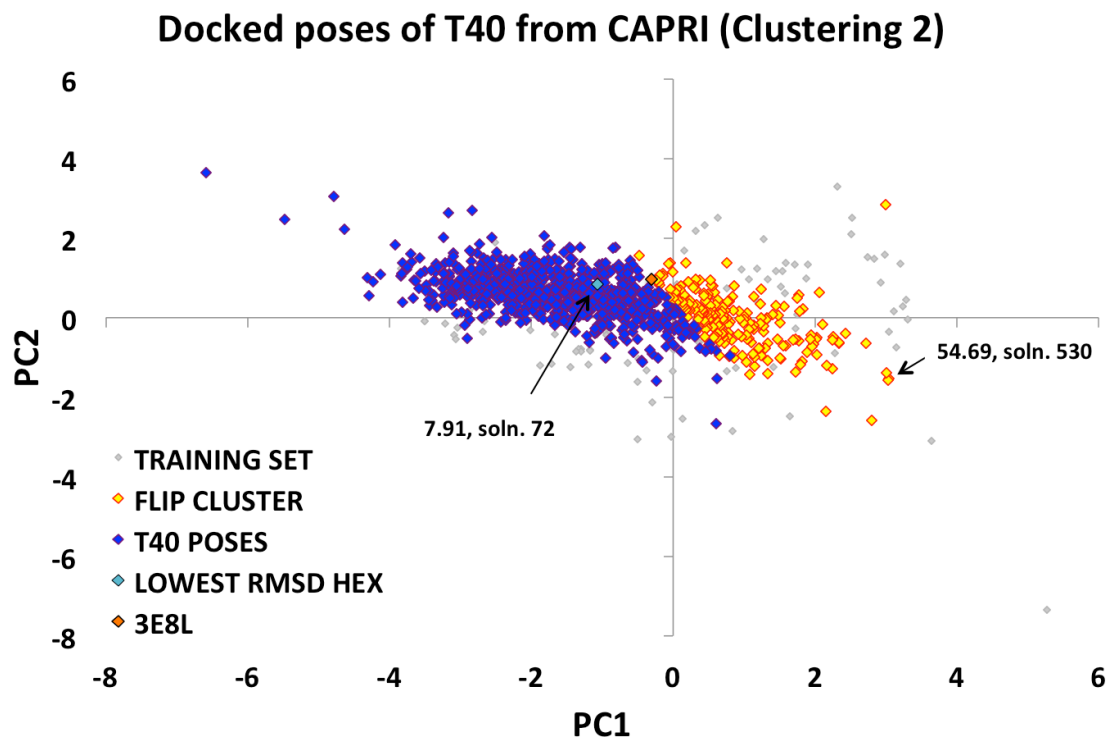
**Figure A1.1.** Docking solution for target 30. a) *2rex* is the target structure. The source structures provided to predict the target were b) *2cls* - chain A and c) *2r2o* - chain B. HEX was used to dock *2cls:A* and *2r2o:B*. 1000 poses were generated and only one structure was identified in the FLIP cluster, d) solution 913 (RMSD 48.31)



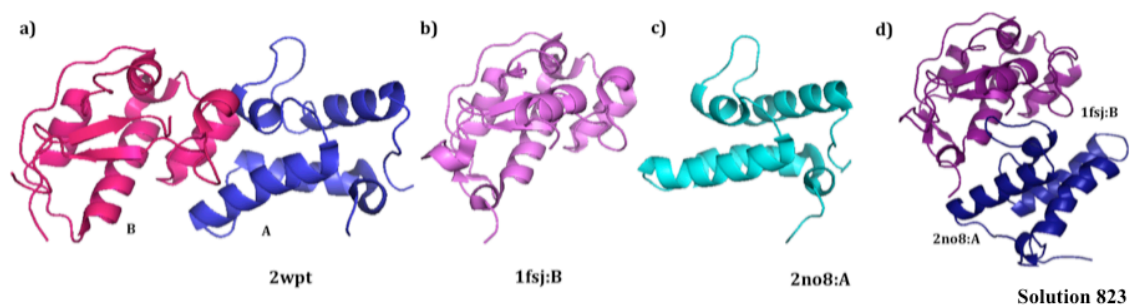
**Figure A1.2.** ECR Analysis of the docking poses of Target 30. Principal Component (PC) values for docking poses are shown plotted against those of the ECR training set (in gray)<sup>13</sup>. Almost 100% poses had PC values in the FunC cluster (blue) while one pose was identified in the FLIP cluster (yellow). The PC values for the structure to be predicted, *2rex*, are plotted in orange and the PC values for the structure with the lowest RMSD generated by HEX is shown in cyan (RMSD 19.78 Å).



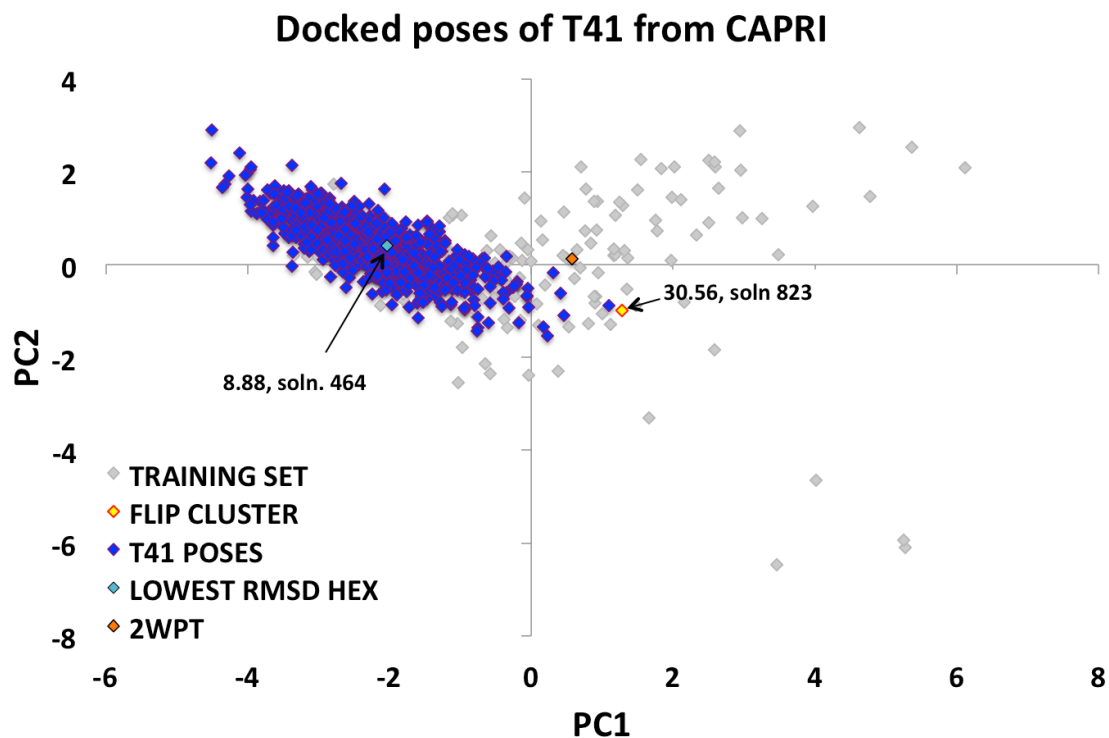
**Figure A1.3.** Docking solution for Target 40. a) *3e8l* has three chains A, B, and C. The target interaction was *3e8l* C:A. The structures used were b) PDBid *3e8l* – chain C, and c) PDBid *1bty* – chain A. d) The solution generated by HEX and identified by ECR as the solution, was pose 530 (RMSD 54.94Å).



**Figure A1.4.** ECR analysis of the docking poses of Target 40. PC values for docking poses are shown plotted against those of the ECR training set (in gray)<sup>13</sup>. 78% of poses were identified in FunC region (blue) and in FLIP region (yellow). The PC values for the structure to be predicted, *3e8l* (C:A) is shown in magenta and the PC values for the structure with the lowest RMSD as generated by HEX is shown in cyan (RMSD 7.91 Å)



**Figure A1.5.** Docking solution for Target 41. a) The target structure *2wpt* has two chains A and B. The structures provided were b) PDBid *1fsj* – chain B and c) PDBid *2no8* – chain A. d) The solution generated by HEX and identified by ECR as the solution was solution 823 (RMSD 30.56Å), shown in Figure 5.5d.



**Figure A1.6.** ECR analysis of the docking poses of Target 41. PC values for docking poses are shown plotted against those of the ECR training set (in gray)<sup>13</sup>. Almost 100% of the poses were in the FunC cluster (blue) while two poses were identified in the FLIP cluster (yellow). The structure to be predicted, *2wpt*, is shown in orange and the structure with the lowest RMSD, as generated by HEX (solution 464), is shown in cyan (RMSD 8.8 Å).