

Multiscale Structural and Biophysical Studies of Protein-Compound Interactions

Stephen Joseph Trudeau

Submitted in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy  
under the Executive Committee  
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2024

© 2024

Stephen Joseph Trudeau

All Rights Reserved

# **Abstract**

Multiscale Structural and Biophysical Studies of Protein-Compound Interactions

Stephen Joseph Trudeau

The recognition of small organic compounds and metabolites is essential for living systems, enabling the cell to sense environmental stimuli and respond appropriately. Developing quantitative models of living systems which can incorporate these environmental stimuli would accordingly benefit from comprehensive mapping of interactions between proteins and small molecules of interest. While high-throughput experimental methods provide a wealth of interaction data, the scale of chemical space currently precludes comprehensive enumeration of protein-compound interaction space. Computational methods can help to bridge this gap by inferring proteome-scale protein-compound interactomes, elucidating structural features within protein families which mediate specificity of binding to specific small molecules, and inferring the affinity of binding for specific protein-compound interactions.

In this thesis, we attempt to use, and in some cases develop, methods to study protein-compound interactions at these three scales. First, we describe recent work in extending our structure-based algorithm for predicting protein-compound interactions throughout the proteome to include a wider array of small molecules. We demonstrate that this method performs comparably to existing methods and describe an online database storing the results of this analysis. We also report several case studies illustrating how this database can be used along with cautionary vignettes indicating areas where the method fails and directions for future improvement. We subsequently analyze druggable pockets occurring within protein-protein interfaces (PPIs) to

assess whether they are less structurally conserved than analogous pockets of conventional drug sites. We find that PPI interfacial pockets are associated with fewer expected off-targets than conventional drug sites, however that this finding is specific to individual protein families, rather than a general feature of interfacial PPI pockets. Finally, we use Free Energy Perturbation to predict the binding affinity of an array of small volatile odorants with an olfactory receptor from the jumping bristletail, *Machilis hrabei*, as well as attempt to further optimize the system in order to study the effects of mutating receptor binding site residues on binding affinity to its active ligands.

# Table of Contents

<i>List of Charts, Graphs, Illustrations</i> .....	<i>iv</i>
<i>Acknowledgments</i> .....	<i>vii</i>
<i>Dedication</i> .....	<i>ix</i>
<b>Chapter 1: Introduction</b> .....	<b>1</b>
<b>1.1 Protein-Compound interactions in Biological Systems</b> .....	<b>1</b>
<b>1.2 Experimental Methods for Characterizing PCIs</b> .....	<b>2</b>
<b>1.3 Computational approaches to Protein-Compound Interaction Prediction</b> .....	<b>4</b>
1.3.1 Ligand-based PCI prediction methods.....	4
1.3.2 Docking and pose scoring methods.....	5
1.3.3 Proteochemometric Methods .....	6
1.3.4 Template-Based PCI Prediction Methods.....	7
<b>1.4 Specific Aims and Thesis Outline</b> .....	<b>8</b>
1.4.1 Specific Aims .....	8
1.4.2 Thesis Outline .....	10
<b>Chapter 2: PrePCI – A structure- and chemical similarity-informed database of predicted protein-compound interactions</b> .....	<b>12</b>
<b>2.1 Introduction</b> .....	<b>12</b>
<b>2.2 Results</b> .....	<b>14</b>
<b>2.3 The PrePCI database – PrePCI/DB</b> .....	<b>22</b>
<b>2.4 Materials and Methods</b> .....	<b>25</b>
<b>2.5 Supplemental Information</b> .....	<b>36</b>

<b>Chapter 3: Applications of PrePCI</b> .....	<b>46</b>
<b>3.1 Introduction</b> .....	<b>46</b>
3.1.1 Additional tools for PCI docking and affinity estimation.....	47
<b>3.2 Results</b> .....	<b>52</b>
3.2.1 Lead compound discovery.....	52
3.2.2 Study of protein-metabolite interactions for protein functional annotation.....	57
3.2.3 Elucidation of drug mechanism of action.....	58
3.2.4 Prediction of Ligands binding at Protein-Protein Interaction Interfaces.....	59
<b>3.3 Limitations of PrePCI structural predictions</b> .....	<b>62</b>
<b>3.4 Discussion</b> .....	<b>65</b>
<b>3.5 Methods</b> .....	<b>67</b>
<b>3.5 Supplemental Information</b> .....	<b>69</b>
<b>Chapter 4: Structural specificity analysis of druggable pockets within protein-protein interaction interfaces</b> .....	<b>88</b>
<b>4.1 Introduction</b> .....	<b>88</b>
<b>4.2 Results</b> .....	<b>95</b>
4.2.1 Druggable PPI pockets are associated with fewer off-targets than active site pockets in some protein families.....	95
<b>4.3 Discussion</b> .....	<b>99</b>
4.3.1 Possible Sources of Bias.....	99
<b>4.4 Materials and Methods</b> .....	<b>102</b>
4.4.1 Materials.....	102
4.4.2 Methods.....	105

4.5 - Supplemental Information .....	108
<b>Chapter 5: Estimation of volatile odorant Binding Affinities to wildtype and mutant MhOR5</b>	
<b>using Free Energy Perturbation .....</b>	<b>109</b>
<b>5.1 Introduction .....</b>	<b>109</b>
5.1.1 Olfactory Receptors .....	109
5.1.2 Biotechnological chemosensors .....	110
5.1.3 Prospective Ligand FEP .....	111
5.1.4 Retrospective Protein FEP .....	112
<b>5.2 Results .....</b>	<b>113</b>
5.2.1 Ligand FEP .....	113
5.2.2 Protein FEP .....	122
<b>5.3 Discussion .....</b>	<b>129</b>
<b>5.4 Materials and Methods .....</b>	<b>130</b>
<b>Chapter 6: Conclusions and Future Directions .....</b>	
<b>134</b>	
<b>6.1 Significance of research .....</b>	<b>134</b>
<b>6.2 Future Directions .....</b>	<b>135</b>
6.2.1 Future Directions for PrePCI .....	135
6.2.2 Alternative approaches to analyzing druggable pockets in PPI interfaces .....	139
6.2.3 Additional analyses of MhOR5 .....	141
<b>Appendix: Code Locations and Information .....</b>	
<b>143</b>	
<b>References .....</b>	<b>145</b>

## List of Charts, Graphs, Illustrations

<i>Table of Contents</i> .....	<i>i</i>
<i>Figure 2.1: PrePCI algorithm for the prediction of protein-compound interactions</i> .....	<i>14</i>
<i>Figure 2.2: PrePCI performance on Pubchem unbiased, all-against-all experimental protein-compound data</i> .....	<i>16</i>
<i>Figure 2.3: Comparison of the performance of PrePCI, LT-scanner (Structure Similarity), and Sequence Similarity.</i> .....	<i>17</i>
<i>Figure 2.4: Performance of PrePCI using different model databases</i> .....	<i>18</i>
<i>Figure 2.5: PrePCI performance on DUD-E and DEKOIS 2.0 PCI datasets</i> .....	<i>19</i>
<i>Table 2.1: Mean difference in per-protein performance between PrePCI and other structure-based PCI prediction methods using the DUD-E and DEKOIS datasets</i> .....	<i>20</i>
<i>Figure 2.6: Comparison of PrePCI performance to other methods on the Directory of Useful Decoys-Enhanced (DUD-E) dataset on a protein-by-protein basis</i> .....	<i>21</i>
<i>Figure 2.7: PrePCI webpage output</i> .....	<i>24</i>
<i>Figure 2.8: Overview of LT-scanner Algorithm</i> .....	<i>29</i>
<i>Figure 2.9: Overview of Morgan Chemical Fingerprint Algorithm</i> .....	<i>31</i>
<i>Figure 2.S1: Performance of PrePCI predictions on unbalanced and balanced PubChem protein-compound interaction experimental data</i> .....	<i>36</i>
<i>Table 2.S1: Comparison of the number of PCIs predicted by LT-scanner and sequence-based metrics at comparable LRs</i> .....	<i>38</i>
<i>Table 2.S2: Comparison of the number of PCIs predicted by LT-scanner using PrePMod and AF/CDD model databases</i> .....	<i>39</i>
<i>Table 2.S3: AUPRC and EF performance of PrePCI, FRAGSITE, FINDSITEcomb2.0 and AutodockVina on the DUDE dataset</i> .....	<i>42</i>
<i>Table 2.S4: EF performance of PrePCI, FRAGSITE, FINDSITEcomb2.0 and vScreenML on the DEKOIS dataset</i> .....	<i>43</i>
<i>Table 2.S5: Composition of the independent drug-target interaction gold standard set<sup>107</sup></i> .....	<i>44</i>

<i>Table 2.S6: Comparison of PrePCI performance to matrix factorization methods on the independent drug-target interaction datasets.</i> .....	45
<i>Figure 3.1 Relative Binding Free Energy Perturbation Thermodynamic Cycle.</i> .....	51
<i>Figure 3.2: Lead compound discovery for ACOT4.</i> .....	54
<i>Figure 3.3: Lead compound discovery for MORC2.</i> .....	56
<i>Figure 3.4 Comparison of template and predicted WWOX-methotrexate binding mode.</i> .....	59
<i>Figure 3.5 PrePCI interaction model of putative BAZ1B-PHIP interaction disruptor, 5XL in complex with PHIP.</i> .	61
<i>Figure 3.6 PrePCI predictions at PPI interfaces.</i> .....	63
<i>Figure 3.7 PrePCI predictions lack metal ions.</i> .....	65
<i>Table 3.S1. PrePCI predicted binders of Ins(1,3,4,5)P4 (PDB id: 4IP).</i> .....	71
<i>Table 3.S2. PrePCI predicted targets of methotrexate and genistein.</i> .....	74
<i>Table 3.S3 Predicted PCIs involving PrePPI predicted master regulator binding proteins.</i> .....	87
<i>Figure 4.1 Pocket-based comparison method for identifying off-targets.</i> .....	94
<i>Figure 4.2 Distribution of expected number of off-targets.</i> .....	97
<i>Figure 4.3: Fraction of off-target proteins as a function of Tversky Index.</i> .....	98
<i>Figure 4.4 Hypothetical Methodological Bias.</i> .....	102
<i>Table 4.S1 Statistical significance of off-target distributions.</i> .....	108
<i>Figure 5.1 Retrospective RBFEP estimation of ligand binding affinity to MhOR5 and benchmarking against experimental affinities.</i> .....	115
<i>Table 5.1 FEP+ <math>\Delta G</math> predictions for retrospective ligand series</i> .....	117
<i>Figure 5.2 Representative poses of methyl laurate, decanal and 2-undecanone bound to MhOR5.</i> .....	118
<i>Table 5.2 FEP+ <math>\Delta G</math> predictions for prospective ligand series</i> .....	121
<i>Figure 5.3 Comparison of compound logP and experimental/predicted binding affinity.</i> .....	121
<i>Figure 5.4 Effect of membrane lipid composition on the ability of FEP+ to predict the effect of MhOR5 mutation on eugenol binding.</i> .....	123
<i>Figure 5.5 Distribution of WT (left) vs mutant (right) MhOR5 contacts with eugenol for M209 and I213 Mutants.</i>	125

*Figure 5.6 Representative pose of eugenol hydrogen bonding to S151 backbone carbonyl in I213A mutant MhOR5.*  
..... 126

*Figure 5.7 RMSD and trajectory analysis of MhOR5 M209 mutants.*..... 128

*Figure 6.1 Possible Application of PrePCI to identify allosteric relationships between ligands for a given protein.*  
..... 138

## Acknowledgments

First and foremost I'd like to thank my thesis advisor, Barry Honig. I am tremendously grateful for the opportunity to work with Barry these past several years and to have been part of his research group. Barry's mentorship has been invaluable in helping me grow as a scientist and learn to think about problems at multiple scales, ranging from the systems level down to molecular biophysics. His careful questioning and discussion served as a constant example of how to think rigorously and systematically, to be skeptical even of positive results, and to find the opportunities among the negative. Later during my time in the lab, Barry provided me with ample room to explore new scientific ideas and computational techniques, encouraging me to seek out gaps in the field which could lead to new, interesting projects and consider all options available to address these unknowns. I am deeply grateful for the care with which Barry approached mentorship and will carry these and other lessons with me throughout my career.

I'd also like to thank several members of the lab, starting with Katie Rosa for all her help in keeping me on track with all manner of logistics throughout grad school. I'd also like to thank Howook Hwang and Deepika Mathur, who laid the foundations of the PrePCI project described in this thesis, as well as Kamrun Begum whose work on the PrePCI website was crucial in making our results accessible in a timely manner. I'd also like to thank Diana Murray, whose attention to detail and openness to brainstorm ideas led us to exciting case studies to highlight the methods we developed, and Donald Petrey who had the answer to any technical question I could imagine. I'd also like to thank Alina Sergeeva and Haiqing Zhao for years of friendship and interesting scientific discussion, batting around ideas for new projects and imagining new directions to explore. It's been a joy working with you all and I look forward to all the exciting projects to come.

I'd also like to thank my thesis committee, Drs. Arthur Palmer, Andrea Califano and Richard Friesner, as well as Dr. Kim Sharp who joined the committee during my defense as an external evaluator. Their advice was invaluable both in developing the research direction throughout grad school and in assembling the final version of this thesis. Additionally, I'd like to thank Dr. Friesner again for his guidance in learning to use FEP and to analyze molecular simulations.

On a more personal note, I'd like to thank all the members of my MD-PhD cohort for being amazing friends and colleagues. A dual-degree program is a long road and there's no group I'd rather walk it with than all of you. I'd like to thank my family, my mother Elizabeth, my father Stephen, my sisters, Jenn and Emily, and of course Emma. Thank you all for a lifetime of support and encouragement; I would have never reached this point were it not for you all. And finally, I'd like to thank my fiancée Tori for her unwavering love and support and for being there at the end of each day.

## **Dedication**

To my family, for instilling in me the curiosity that led me to research

# Chapter 1: Introduction

## 1.1 Protein-Compound interactions in Biological Systems

The precise recognition and interaction of biomolecules is fundamental to the coordination of biological processes and enabling the cell to respond to environmental stimuli. On the macromolecular scale, complex networks of protein-protein interactions mediate the cell's internal processes while tightly regulated interactions between proteins and nucleic acids control the cell's transcriptional state<sup>1</sup>. Advances in high-throughput experimental methodologies have dramatically accelerated the process of cataloging protein-protein and protein-DNA interactomes in a wide range of organisms and experimental conditions. Yeast-two-hybrid<sup>2-4</sup> and mass spectrometry<sup>5, 6</sup> methodologies have identified hundreds of thousands of protein-protein interactions (PPIs) while DNA-microarrays<sup>7</sup> and CHIP-seq<sup>8, 9</sup> have enabled genome-scale detection of protein-DNA interactions. Combined with additional context-specific data sources capturing transcriptional states<sup>10-12</sup> and post-transcriptional modifications<sup>13</sup>, these interactions have enabled myriad computational strategies to unravel the systems used by cells to survive and adapt<sup>14-24</sup>.

While elucidation of these macromolecular interactions allows us to identify the modules by which cells respond to stimuli, the stimuli themselves are often in the form of environmental small molecules and metabolites. Cells typically detect the presence of these small molecules via their interactions with proteins, simultaneously integrating numerous signals and directing the cell to respond accordingly<sup>25, 26</sup>. Such protein compound interactions (PCIs) act at multiple scales, playing key roles controlling metabolism and gene expression, directing proteins to their proper subcellular environments and detecting changes in the external environment. For example,

feedback inhibition of phosphofructokinase, a central glycolytic enzyme, is classically achieved through allosterically binding its downstream product, adenosine triphosphate<sup>27</sup>. Expression of lac operon genes associated with lactose utilization are regulated by proteins which interact with glucose, the cell's preferred energy resource, and lactose, an alternative sugar used when glucose is lacking<sup>28</sup>. And peripheral membrane proteins, which recognize the precise phosphorylation patterns of membrane phospholipid headgroups, are recruited in response to external stimuli to the appropriate cellular surfaces<sup>29,30</sup>. Given the numerous ways by which protein recognition of small molecules can modify cellular activity, inclusion of protein-compound interaction data into systems-level applications may facilitate the development of more realistic and quantitative models.

## **1.2 Experimental Methods for Characterizing PCIs**

To effectively capture the influence of PCIs on cellular networks, comprehensive data characterizing PCIs at both proteomic and metabolomic scales are needed. To that end, numerous experimental methods have been developed to characterize the structural, thermodynamic and network characteristics of PCIs. Structural approaches such as X-ray crystallography<sup>31</sup>, cryo-electron microscopy<sup>32,33</sup> and nuclear magnetic resonance<sup>34</sup>, though among the lowest throughput of experimental modalities, yield detailed structural models of PCIs which can reveal the physical basis of small molecule recognition and specificity. Quantitative affinity techniques such as isothermal titration calorimetry<sup>35</sup> and surface plasmon resonance<sup>36</sup> provide precise measurements of interaction affinities, characterizing the effective concentration range over which ligand binding is likely to be significant. Additionally, recently developed mass-spectrometry methods have greatly facilitated the characterization of PCI networks. Protein-centric methods such as MIDAS

simultaneously identify multiple metabolites which concentrate within a chamber containing a protein of interest relative to a control well separated by a semi-permeable membrane<sup>26, 37</sup>. In contrast, compound-centric approaches enable proteome-scale detection of proteins which interact with a compound of interest. Drug affinity responsiveness target stability (DARTS)<sup>38</sup> and Limited-Proteolysis Small Molecule Mapping (LiP-SMap)<sup>25, 39-41</sup> use partial proteolysis of lysates from cells incubated with a test compound. Binding of a test compound sterically protects protease cleavage sites, resulting in novel peptides from ligand-bound proteins which can be used to identify the ligand binding sites proteome-wide. Similarly Stability of Proteins from Rate of Oxidation (SPROX)<sup>42-44</sup> identifies methionine residues which are protected from hydrogen peroxide mediated oxidation in the presence of a small molecule. From these protected sites and the “conformotypic peptides” obtained from LiP-SMap, it is possible to infer both which proteins bind to a small molecule of interest as well as the likely binding site<sup>25, 40, 42-44</sup>.

These and related methods have uncovered numerous previously unknown PCIs and suggest multiple mechanisms for cross-talk between metabolic pathways. Hicks et al. used MIDAS to identify 830 interactions between 33 enzymes involved in carbohydrate metabolism and a panel of 401 metabolites. In addition to uncovering cell-type specific regulation of lactate dehydrogenase by ATP, they found numerous interpathway protein-metabolite interactions involving key metabolic enzymes such as glucokinase, phosphofructokinase and isocitrate dehydrogenase<sup>26</sup>. Piazza et al. used LiP-SMap to identify over 1600 interactions involving 20 central carbon metabolites in the E. coli proteome, of which less than 14% were previously known<sup>25</sup>. Moreover, they found that many PCIs occurred in proteins whose expression levels were relatively constant, whereas variably expressed proteins interacted with comparatively few metabolites, prompting them to suggest that transcriptional regulation of protein activity may be independent or even

mutually exclusive of metabolite-mediated regulation<sup>25</sup>. Results such as these hint at how incomplete our understanding of protein-compound interaction networks are and how much of protein-compound interaction space remains to be explored.

### **1.3 Computational approaches to Protein-Compound Interaction Prediction**

Although the growth in experimental methods has generated a plethora of data, with drug-like chemical space estimated to contain more than  $10^{60}$  compounds, even state-of-the-art experimental approaches can sample only a tiny fraction of chemical space. Computational methods, while still limited in scope to billions of compounds, have the promise of filtering through chemical space more rapidly to prioritize PCIs for experimental validation. Traditional computational methodologies tend to fall largely into four broad categories, ligand-based, docking-based, proteochemometric, and template-based with machine-learning and deep-learning methodologies increasingly applied to each of these domains.

#### **1.3.1 Ligand-based PCI prediction methods**

Ligand-based methods infer novel PCIs for a protein of interest by identifying compounds that are chemically similar to known binding ligands<sup>45</sup>. Approaches based on quantitative structure activity relationships (QSAR) attempt to identify common physical and topological properties of compounds that mediate interaction with a protein of interest<sup>46</sup>. Additionally, analysis of ligand binding ensembles has been used to identify relationships between proteins based on their shared pharmacology<sup>47</sup>. Such approaches have benefitted greatly from the development of machine learning methods ranging from Linear Regression Models<sup>48</sup>, Support Vector Machines<sup>49</sup>, Bayesian Neural Networks<sup>50, 51</sup>, Random Forest<sup>52</sup>, Self-Organizing Maps<sup>53</sup>, Convolutional Neural Networks

(CNNs)<sup>54</sup> and Graph Neural Networks (GNNs)<sup>55</sup> which enable automatic identification and weighting of the features which can most reliably be used to reproduce the training data.

Ligand-based approaches are typically computationally efficient and scale well with chemical library size<sup>45</sup>. However, because they typically require a rich dataset of ligands that are known to interact with a protein of interest, their utility is often limited to proteins with existing chemical interaction data. As most proteins in the human proteome lack such data, integration with other methodologies, such as those described in section 1.3.4 and Chapter 2 of this thesis, are required to apply ligand-based methods proteome-wide.

### **1.3.2 Docking and pose scoring methods**

Docking-based approaches sample and score protein-compound interaction poses to provide a quantitative estimate of binding affinity<sup>56-61</sup>. Application of such approaches to new proteins has historically been limited by the need for a high-quality structure, ideally an experimental holo-structure of the protein bound to another compound, as performance on protein apo-structures suffers from not accounting for protein structural rearrangements which occur on ligand binding<sup>58</sup>. However, this restriction has been somewhat alleviated by the recent development of methods which generate ensembles of protein structures to simulate protein flexibility and prioritize poses that are more conducive to ligand binding<sup>58, 62</sup>. Traditionally, scoring functions based on physical forcefields have been used to score the sampled poses<sup>56, 57, 60</sup>, however machine learned scoring functions based on random forests<sup>63-65</sup> or neural networks<sup>66-68</sup> trained on binding site interatomic distances, and more recently, CNNs<sup>69-72</sup> and GNNs<sup>73, 74</sup> trained on PCI holo-structures directly, have been applied to infer binding affinities. Moreover, fingerprint methods, which identify interaction motifs and score putative interaction poses based on their similarity to a structural template, have shown great promise<sup>75-77</sup>. Such docking-based strategies

have enabled structure-based virtual screening of hundreds of millions to billions of small molecules against specific protein targets of interest and have discovered novel protein chemotypes<sup>78, 79</sup>. However, the computational costs of pose generation and scoring currently prevent docking methods from being applied at a proteome-scale.

### 1.3.3 Proteochemometric Methods

In contrast to ligand and docking-based methods, proteochemometric methods infer PCIs using independent protein and compound features. Algorithms such as REMAP<sup>80</sup>, COSINE<sup>81</sup>, NRLMF<sup>82</sup>, and MDMF2A<sup>83</sup>, formulate PCI prediction as a matrix factorization problem in which low-rank matrices representing abstract protein and chemical features are derived from protein sequence similarity and chemical similarity, respectively<sup>80</sup>. All pairwise protein-compound pairs can be efficiently scored by multiplying these low-rank matrices. 3D-REMAP augments REMAP with ligand binding site similarity and binding affinities for compounds of interest<sup>84</sup>, however the inclusion of this docking step limits the wider applicability to proteome-scale applications. Additional methods based on Bayesian Additive Regression Trees<sup>85</sup>, Support Vector Machines<sup>86, 87</sup>, Random Forests<sup>87</sup>, CNNs<sup>88, 89</sup> and large-language model embeddings of protein and compound features<sup>90</sup> have also been developed.

Because proteochemometric methods do not require pose generation for each PCI, they are potentially amenable to proteome-scale PCI prediction. Moreover, the embedding functions they use could, in principle, enable them to effectively encode and score large regions of chemical space. However, the transferability of these methods to proteins or compounds outside of their training sets remains unclear. Chen et al. found that performance of deep-learning proteochemometric models trained on ligands alone was statistically indistinguishable from models which included protein features<sup>89</sup>. They further noted that model weights for protein

features were tightly distributed around zero, whereas the ligand weights were more widely dispersed, suggesting that current methods are not effectively learning from protein features, rather exploiting hidden biases within the ligand dataset and are therefore unlikely to transfer to new proteins<sup>89</sup>.

### **1.3.4 Template-Based PCI Prediction Methods**

In contrast, template-based approaches analyze protein surfaces and compare them to the structures of known compound binding sites. Such methods, though they do not capture the physics of protein-compound interactions with the fidelity of docking-based approaches, can rapidly scan databases of protein structures to identify surface regions similar to known binding sites in order to prioritize and orient subsequent docking and experimental follow-up. Unlike docking and ligand-based methods, template-based methods can readily scale to proteome-wide applications. Moreover, unlike proteochemometric methods which tend to characterize proteins via their primary structure or 2D contact maps, template-based methods fully utilize protein structural similarity, allowing for the detection of remote predictions that are not obvious by sequence while preserving interpretability and transferability.

Our lab recently developed LBias and LT-scanner<sup>91</sup>, related template-based methods that identify novel protein-compound interactions by superimposing homology models onto experimental structures of PCIs taken from the Protein Data Bank (PDB). The similarity between the template and the model ligand binding sites is quantified using a custom metric called a SIM score which describes the degree to which the query could recapitulate the interactions made by the template protein with the compound. The FINDSITE suite of methods developed by the Skolnick group takes a similar approach<sup>92, 93</sup>, using threading to generate structural models of query proteins based on ligand-bound template structures and identifying the ligands bound to

evolutionarily conserved sites. However, whereas LT-scanner scores the similarity of the analogous sites on the query and target structures, FINDSITE uses chemical similarity of ligands identified in template binding sites as the basis of its scoring function.

## **1.4 Specific Aims and Thesis Outline**

### **1.4.1 Specific Aims**

In this thesis, we aim to develop and apply methods to study protein-compound interactions at numerous scales. At the proteome-scale, we incorporate chemoinformatic measures of chemical similarity to extend LT-scanner to larger regions of chemical space. At the family level, we perform quantitative comparisons of protein surface pockets to assess whether PPI interfacial pockets are associated with fewer off-targets than conventional drug pockets. And finally at the individual PCI level, we use Free Energy Perturbation to study the ligand binding preferences of an insect olfactory receptor.

#### **1.4.1.1 Extension of proteome-scale PCI prediction method**

Our lab has previously developed LT-scanner<sup>91</sup>, a structure-based method for identifying novel PCIs based on structural similarity between protein models and template protein-compound holostructures in the PDB<sup>94</sup>. While LT-scanner was shown to be effective in retrospective benchmarking, its reliance on template holo-structures restricts its applicability to compounds which are already co-crystallized with at least one protein within the PDB. In this work, we aim to extend LT-scanner by integrating its predictions with chemical similarity metrics to increase the range of compounds which can be considered from the roughly 30,000 compounds in the PDB to over 6.8 million compounds. We perform several benchmarking studies to evaluate the reliability

of our predictions, provide several case studies of possible uses of this updated method and note cases where PrePCI provides seemingly implausible predictions which could be the basis of future work.

#### **1.4.1.2 Family level evaluation of the specificity of druggable binding sites within protein-protein interfaces**

A recent analysis of the PDB identified over 160,000 pockets near and within PPI interfaces which are predicted to be druggable<sup>95</sup>. While the authors compared these pockets to those of conventional inhibitors, they did not indicate whether PPI pockets are any more structurally unique than conventional drug sites. If PPI pockets are indeed more structurally unique than conventional drug sites, drugs targeting PPI pockets may be able to do so more selectively with fewer off-targets and side-effects. We propose and apply a method for studying the similarity of PPI interfacial pockets within protein families to assess whether ligands targeting PPI pockets are likely to affect fewer off-target proteins than conventional active site drugs. We subsequently address possible shortcomings of the proposed approach and provide suggestions on improving the analytical method.

#### **1.4.1.3 Free Energy Perturbation analysis of the *Machilis hrabei* olfactory receptor, MhOR5**

Computational prediction of binding free energies is a long-standing goal of computational chemistry. In this thesis, we use free energy perturbation (FEP) to study the binding preferences of MhOR5, an olfactory receptor of the jumping bristletail, *Machilis hrabei*. First, we attempt to prospectively predict MhOR5's binding affinities with a panel of small molecules associated with tuberculosis, malaria and Sars-CoV-2 infection. Secondly, we evaluate the impact of membrane lipid composition on the accuracy of protein mutation FEP in an effort to optimize a model system for additional prospective simulations.

## 1.4.2 Thesis Outline

This thesis is organized as follows. In Chapter 2, we present the materials and methods underlying PrePCI, our structure-based algorithm for predicting protein-compound interactions at a proteome scale, as well as benchmarking analyses demonstrating its accuracy and comparing to other high-throughput computational methods. Building on our lab's previous method, LT-scanner<sup>91</sup>, we implement a chemoinformatic screen to identify novel compounds which are not present in the PDB but are similar to PDB compounds, and integrate chemical similarity scores with structure and sequence similarity scores using a Naïve Bayes framework called PrePCI. In chapter 3, we describe several case studies applying PrePCI to identify lead compounds for drug discovery, elucidate a potentially unrecognized drug mechanism of action and suggest a novel biological function via the prediction of protein-metabolite interactions. We also describe several case studies in which PrePCI provides implausible predictions to alert the user to possible pitfalls as well as suggest potential avenues for future development. In chapter 4, we analyze the structural uniqueness of druggable pockets within PPI interfaces of 6 protein families and estimate the number of off-target proteins which would likely be affected by drugs targeting these pockets. We compare the number of off-targets expected for these PPI pockets to the number of off-targets expected from conventional active site pockets. We find that druggable pockets in PPIs are associated with fewer expected off-targets than conventional active site pockets for 3 of the 6 families, though we also note the several limitations of the data and methodological approach and discuss means of ameliorating these limitations in future work. In chapter 5, we perform more detailed biophysical studies of a specific protein system, the *Machilis hrabei* olfactory receptor 5 (MhOR5), performing free energy perturbation studies to prospectively predict the binding of MhOR5 to an array of volatile odorants. We also test whether using different membrane lipids

improves retrospective free energy perturbation analysis of MhOR5 binding site mutations. Finally in chapter 6, we discuss limitations of the current work and discuss possible directions to proceed in the future.

## **Chapter 2: PrePCI – A structure- and chemical similarity-informed database of predicted protein-compound interactions**

The following two chapters are adapted from:

Trudeau, S.J., Hwang, H., Mathur, D., Begum, K., Petrey, D., Murray, D., and Honig, B. (2023). PrePCI: A structure- and chemical similarity-informed database of predicted protein compound interactions. *Protein Sci* 32, e4594. 10.1002/pro.4594.<sup>96</sup>

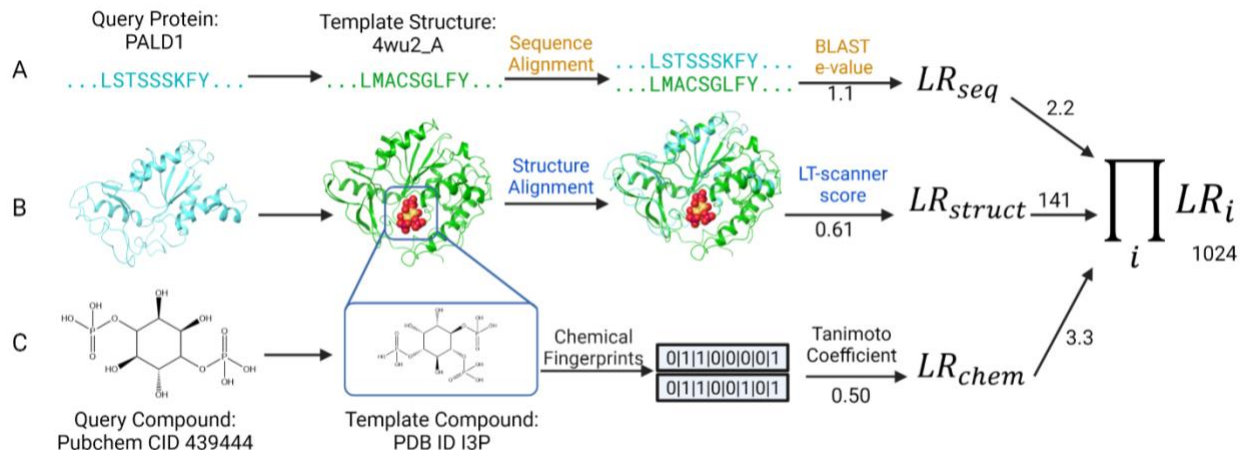
### **2.1 Introduction**

Over the following two chapters, we describe the development, evaluation and applications of PrePCI, an extension of our lab's structure-based method for predicting protein-compound interactions at a proteome-scale. In this chapter, we review the resources and methods used in the development of PrePCI which expands our coverage of chemical space by over 200-fold. We then proceed to benchmark PrePCI in an unbiased manner against high-throughput PCI data obtained from the Pubchem bioactivity database, as well as on smaller, more curated datasets. PrePCI's performance relative to other methods is evaluated on these datasets. Finally, the PrePCI database and web-application are described.

#### **2.1.1 Overview of PrePCI Algorithm**

The PrePCI algorithm is depicted in Figure 2.1 and consists of three components. Figure 2.1A illustrates the sequence similarity component where a query protein sequence is matched to protein sequences from PDB template-compound complexes using BLAST. A query protein is predicted as a target of a compound which appears in a PDB complex based on the sequence alignment score of the query with the template protein (see Methods). Figure 2.1B illustrates the

structural component, mediated by the program LT-scanner<sup>91</sup>, in which a query protein is structurally aligned to a template complex in the PDB<sup>94</sup> with the Ska program<sup>97</sup>. The transformation that aligns the two proteins is used to place the PDB compound in the coordinate frame of the query protein. The LT-scanner scoring function<sup>91</sup> assesses the compatibility of the compound with the query protein by calculating a score based on the extent to which residues in the query binding site recapitulate the physicochemical interactions between the protein and compound in the template complex. The sequence similarity and LT-scanner calculations are performed to compare all query protein sequences and models against all PDB protein-compound templates. Figure 2.1C illustrates the chemical similarity component where PDB compounds are matched to topologically similar compounds from the PubChem database. Both PDB and Pubchem compounds are converted to molecular fingerprints representing the submolecular fragments present in each molecule<sup>98</sup>. These fingerprints are then quantitatively compared using the Tanimoto Coefficient (TC), or Jaccard Index, which represents the number of fragments the two compounds have in common, normalized by the total number of distinct fragments represented by the two compounds, i.e. the size of the intersection of the two sets normalized by the size of their union<sup>99</sup>. When the TC between a PDB compound and a Pubchem compound exceeds 0.5, the Pubchem compound is predicted to target the protein found by either sequence similarity (Figure 2.1A) or LT-scanner (Figure 2.1B). A Bayesian procedure is used to integrate the sequence, structure, and chemical similarity scores for each query protein-compound prediction into a likelihood ratio derived from a true positive set of experimentally characterized PCIs. The scored PCI predictions comprise the PrePCI database (PrePCI/DB).



**Figure 2.1: PrePCI algorithm for the prediction of protein-compound interactions.**

PrePCI uses BLAST and LT-scanner for the query protein (aqua sequence and model, left) to identify proteins within PDB protein-compound complexes that have (A) sequence and/or (B) structure similarity to the query. (C) Compounds predicted to bind the query are identified using a Tanimoto coefficient (TC) chemical similarity search of fingerprints representing the PDB compound and compounds from PubChem. In this example, the sequence and model for Paladin (PALD1) are matched to the PDB complex (PDB ID 4wu2) of *Selenomonas ruminatum* myo-inositol hexaphosphate phosphohydrolase bound to the PDB compound I3P with BLAST e-value 1.1 and LT-scanner score 0.61. A query compound from PubChem (CID 439444) has TC = 0.5 with the PDB compound. The LR for the interaction between PALD1 and the query compound is the product of the LRs from sequence, structure, and chemical similarity scores.

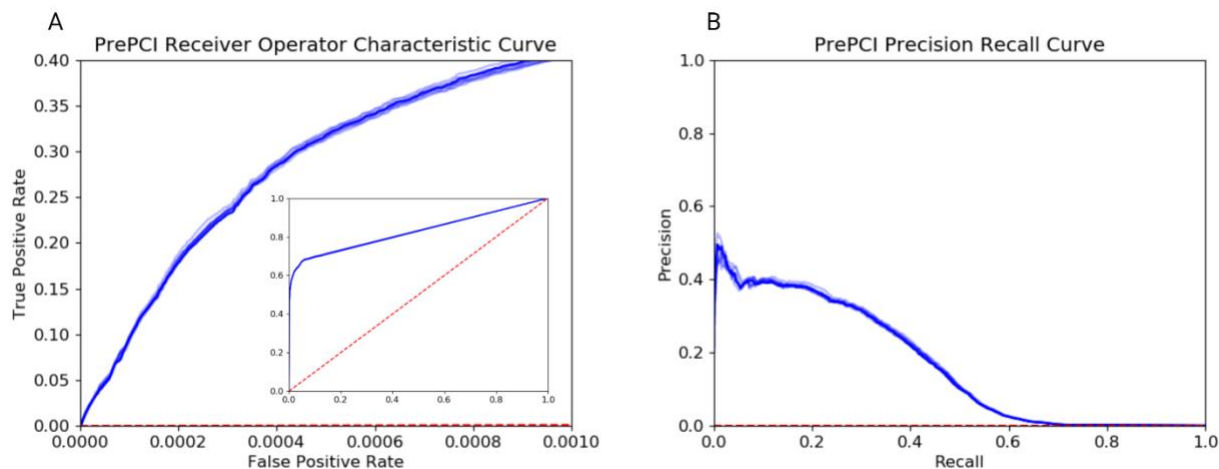
## 2.2 Results

### PrePCI Training and Evaluation

To evaluate PrePCI's performance, a Naïve Bayes Classifier was trained using 10-fold cross-validation using a true positive set of PCIs with bioactivity data from PubChem<sup>100, 101</sup>. As described in Methods, the true positive set consists of 285K PCIs for 142K compounds and 2,926 proteins. The negative set consists of 417M hypothetical PCIs between the 142K compounds and 2,926 proteins for which PubChem provides no bioactivity data. For each of the ten folds, PrePCI's performance was evaluated by ranking predictions by their overall likelihood ratio (LR) and computing the area under the receiver operator characteristic curve (AUROC) and the average

precision (or area under the precision-recall curve, AUPRC) (Figure 2.2). The resulting ROC and Precision-Recall curves are highly concordant, with mean AUROC and average precision of  $0.828 \pm 0.001$  and  $0.165 \pm 0.002$ , respectively. It is important to note that, due to the size of the negative set, the testing set is heavily imbalanced with negatives outnumbering positives by approximately 1400:1, corresponding to a random precision of  $7 \times 10^{-4}$ . The average precision of 0.165 therefore constitutes a substantial enrichment of true positive predictions and is likely an underestimate as many PCIs considered false positives presumably correspond to as yet unannotated true interactions.

Moreover, in reviewing the experimentally known PCIs with low PrePCI scores corresponding to the high false positive region of the ROC curve, we found these PCIs are primarily cases where PrePCI could not identify a template compound that was both similar to the query compound and predicted to bind the query protein. Consequently, these PCIs could not be effectively scored, limiting the maximum AUROC we could obtain, as reflected in the sharp elbow in the ROC curve (Figure 2.2A). The region to the right of this elbow corresponds to a linear interpolation from the TPR/FPR associated with the least confident PrePCI prediction to the point (1,1), rather than assessing PCIs which were meaningfully scored. To evaluate PrePCI's performance on its meaningful predictions, we recomputed ROC and Precision-Recall curves for each of the 10 folds, restricting the evaluation to PCIs where a template could be identified (see Methods). This restriction resulted in a true positive set composed of 204,919 PCIs and a true negative set of 62,414,150 PCIs, constituting 72% and 15% of the original true positive and true negative sets respectively. Evaluating only scored interactions yielded improved performance in AUROC and AUPRC to  $0.936 \pm 0.001$  and  $0.235 \pm 0.002$  with a more constant precision over most of the recall range (See Figure 2.S1 and related discussion).



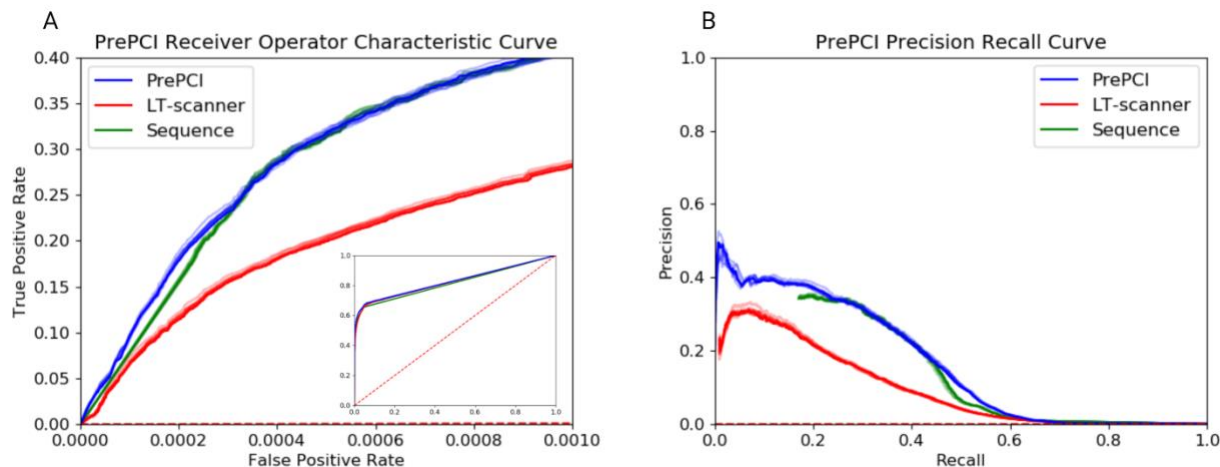
**Figure 2.2: PrePCI performance on Pubchem unbiased, all-against-all experimental protein-compound data.**

(A) Receiver operating characteristic curve and (B) Precision Recall curve for each of the 10 folds of cross-validation for training and testing PrePCI on experimentally observed PCIs from PubChem. Curves corresponding to the data fold which yielded the median area under the ROC curve (AUROC, A) and average precision (B) across all cross-validation folds are darker, while curves for remaining folds (lighter blue) are included to display the range of results obtained from the individual folds. PrePCI's average AUROC and Average Precision on the PubChem dataset are  $0.828 \pm 0.001$  and  $0.165 \pm 0.002$  respectively.

### LT-Scanner and Sequence Similarity are Synergistic

PrePCI performance was compared to the performance of classifiers using features based only on sequence similarity or LT-scanner alone. Sequence similarity outperforms LT-scanner and performs comparably to PrePCI, although PrePCI's use of both enables superior performance (Figure 2.3). In ROC curve analysis, PCIs present in the PDB were excluded from LT-scanner testing, however a sequence identity cutoff was not implemented, therefore many of the sequence similarity targets are likely to be obvious and, thus, underlie the performance of the sequence similarity classifier. The unique feature of LT-scanner is its ability to identify non-trivial relationships. Indeed, as can be seen in Table 2.S1, LT-scanner identifies many more relationships than are available from sequence similarity alone. The combined use of sequence and structure

yields the greatest coverage of true positive PCIs without impairing performance as each method identifies PCIs not detected by the other at comparable LRs (Figure 2.3, Table 2.S1).



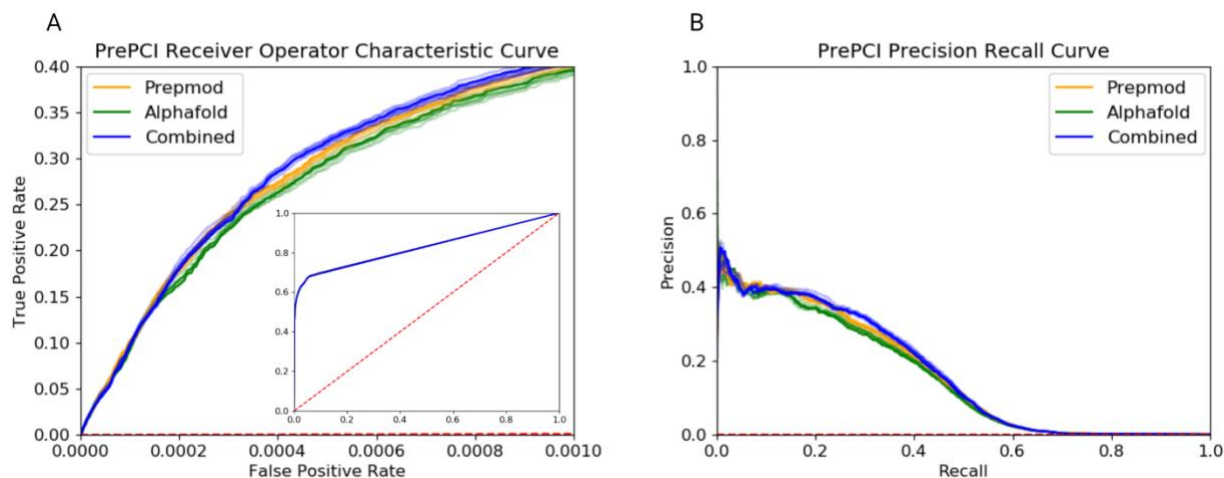
**Figure 2.3: Comparison of the performance of PrePCI, LT-scanner (Structure Similarity), and Sequence Similarity.**

(A) ROC and (B) Precision-Recall Curves comparing PrePCI (blue) to the performance obtained using LT-scanner<sup>102</sup> and sequence similarity (green) alone, where all pairwise PCIs within the PubChem Benchmarking set are ranked by the LT-scanner score and sequence similarity score. The average AUROC for PrePCI, LT-scanner and Sequence similarity are  $0.828 \pm 0.001$ ,  $0.825 \pm 0.001$  and  $0.816 \pm 0.002$ , respectively. The mean Average Precision for PrePCI, LT-scanner and Sequence similarity are  $0.165 \pm 0.002$ ,  $0.095 \pm 0.001$  and  $0.150 \pm 0.002$ , respectively.

### The Union of Homology Models and AlphaFold Structures as Targets Increases PCI Coverage

PrePCI performance was evaluated with predictions for query structures from PrepMod, our lab's in-house homology model database, versus a domain-level database of protein models extracted from AlphaFold 2.0<sup>102, 103</sup> (AF/CDD, see Methods). As shown in Figure 2.4, performance is similar regardless of the query model database used. While the number of predictions is greater with AF/CDD versus PrepMod structures, the combination of the databases is synergistic and results in the highest number of PCI predictions (Table 2.S2) and best performance on both ROC and Precision-Recall Curve analysis (Figure 2.4). For example, as depicted in the first row of Table 2.S2, in cases where the query model aligns well with the template complex (LT-scanner score  $\geq$

0.6), PrePCI-PrePMod predicts 64K PCIs and PrePCI-AF/CDD predicts 77K PCIs. The intersection of the two sets is 39K PCIs and the union is 101K PCIs.



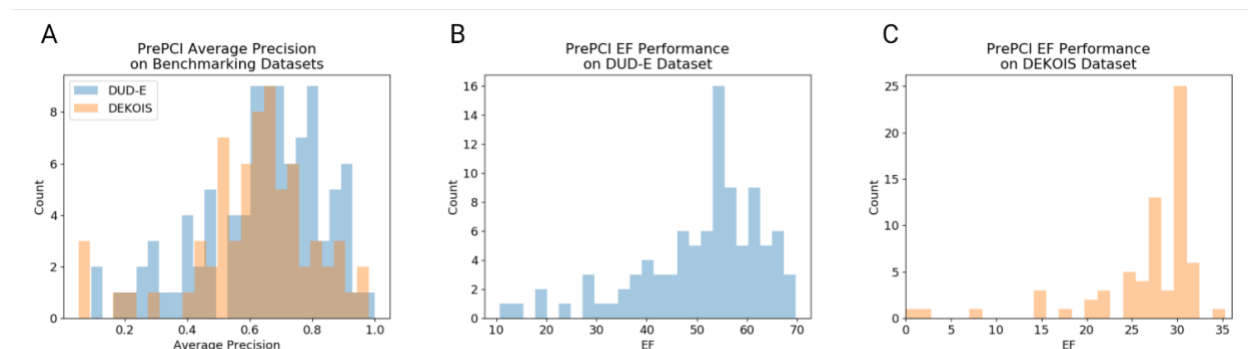
**Figure 2.4: Performance of PrePCI using different model databases.**

(A) ROC and (B) Precision-Recall curves obtained by 10-fold cross-validation using all pairwise PCIs within the PubChem Benchmarking set when PrePCI is restricted to structural predictions made using AF/CDD (green) or PrePMod alone (orange) as the model database. For comparison, the results from Figure 2 in which the highest LT-scanner score, irrespective of whether it was based on a PrePMod or AF/CDD model (blue), are included. The average AUROC obtained when using AF/CDD, PrePMod, and the combination are  $0.828 \pm 0.001$ ,  $0.828 \pm 0.001$  and  $0.828 \pm 0.001$  respectively and the mean Average Precision were  $0.163 \pm 0.002$ ,  $0.156 \pm 0.002$  and  $0.165 \pm 0.002$  respectively.

#### PrePCI performance on Independent Docking Datasets

To compare PrePCI to other structure-based PCI prediction methods, we benchmarked its performance on two commonly used docking benchmarks, the Directory of Useful Decoys-Enhanced (DUD-E)<sup>104</sup> and the Demanding Evaluation Kits for Objective In Silico Screening 2.0 (DEKOIS 2.0)<sup>105</sup>. These datasets consist of proteins with associated active compounds which are known to bind the protein, and property matched decoy small molecules which do not bind the protein. Using leave-one-out cross-validation, PrePCI achieves a mean average precision of  $0.64 \pm 0.20$  and a mean enrichment factor,  $EF_{0.01}$ , of  $51.0 \pm 12.5$  across the 95 targets in DUD-E

(Figure 2.5A,B, Table 2.S3), while achieving a mean average precision of  $0.62 \pm 0.19$  and a mean enrichment factor,  $EF_{0.01}$ , of  $26.6 \pm 6.5$  on across the 69 human targets in DEKOIS 2.0 (Figure 2.5A,C Table 2.S4).



**Figure 2.5: PrePCI performance on DUD-E and DEKOIS 2.0 PCI datasets.**

Histograms indicating the distribution of (A) per-protein average precision for the DUD-E (blue) and DEKOIS 2.0 (orange) datasets and the distribution of  $EF_{0.01}$  for the DUD-E (B) and DEKOIS 2.0 (C) datasets.

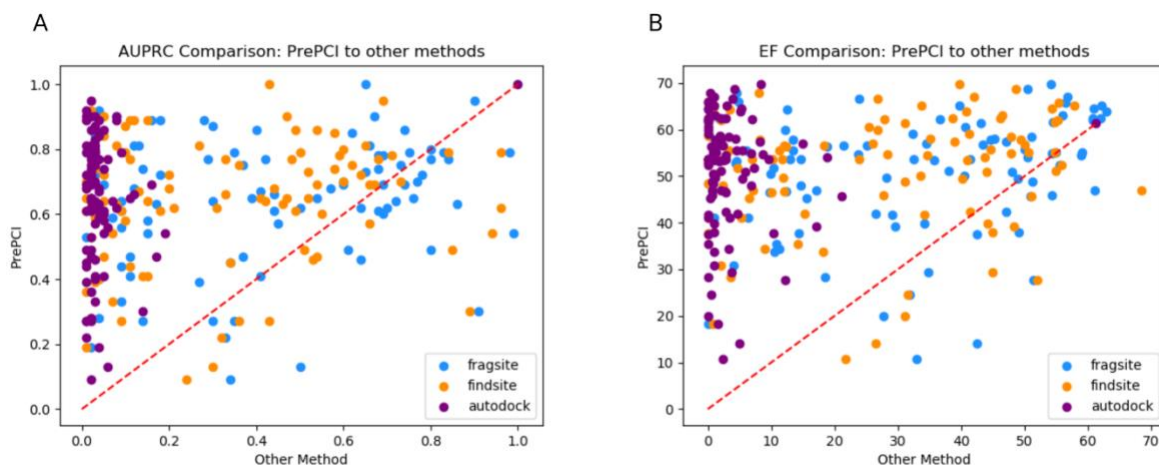
We subsequently compared the per-protein performance of PrePCI to FRAGSITE<sup>93</sup>, FINDSITE<sup>comb2.0 92</sup> and Autodock Vina<sup>60</sup> for all 95 human proteins in DUD-E using the Wilcoxon Signed Rank Test (Figure 2.6A, 2.6B, Table 2.S3). After correcting for multiple comparisons using a Bonferroni correction, we found that PrePCI significantly outperformed all three of these other methods (Table 2.1). Additionally, we trained PrePCI using the PCI data aggregated from 53 human DEKOIS proteins, consistent with previous benchmarking studies for vScreenML<sup>106</sup> and FRAGSITE<sup>93</sup>, and individually tested PrePCI on each of the remaining 20 human proteins in the DEKOIS 2.0 dataset. While PrePCI significantly outperformed vScreenML, performance was comparable to that of FRAGSITE and FINDSITE<sup>comb2.0</sup> after Bonferroni Correction (Table 2.1).

	<i>DUD-E EF</i>	<i>DUD-E AUPRC</i>	<i>DEKOIS EF</i>
<i>FragSite</i>	20.07 (3.52E-13)	0.24 (2.51E-9)	7.68 (0.24)
<i>FindSite</i>	24.96 (4.98E-18)	0.31 (1.79E-14)	8.45 (0.15)
<i>Autodock</i>	47.25 (2.34E-16)	0.59 (3.41E-16)	(-)
<i>vScreenML</i>	(-)	(-)	16.00 (9.00E-4)

**Table 2.1: Mean difference in per-protein performance between PrePCI and other structure-based PCI prediction methods using the DUD-E and DEKOIS datasets.**

For each protein from the test set indicated by the column label, the difference between PrePCI’s performance and that of FragSite, FindSite, Autodock and vScreenML was computed as described in Methods. The average difference in performance is reported along with the Bonferroni adjusted p-value (defined as  $\min(1, 9 * p\text{-value})$ ) obtained using a Wilcoxon Signed Rank Test.

It is important to note that the results reported for both FRAGSITE and FINDSITE<sup>comb2.0</sup> were obtained using a template sequence identity cutoff of 30%, and therefore may not represent the optimal performance if these methods were permitted to use all available templates, including those with comparatively high sequence similarity. Because PrePCI relies on pre-computed homology models which are constructed using the best template available, we were unable to benchmark PrePCI using a similar sequence cutoff and therefore tentatively conclude that PrePCI likely performs similarly to FRAGSITE and FINDSITE. However, we emphasize that the size and accessibility of the PrePCI database renders it a particularly unique and comprehensive resource.



**Figure 2.6: Comparison of PrePCI performance to other methods on the Directory of Useful Decoys-Enhanced (DUD-E) dataset on a protein-by-protein basis.**

(A) PrePCI tends to outperform FRAGSITE, FINDSITE and AutoDock Vina for most proteins in the DUDE dataset when evaluated using the Area Under the Precision Recall curve. PrePCI's outperformance is statistically significant using a Wilcoxon Signed Rank Test. (B) PrePCI's outperformance is similarly significant when evaluated using the Enrichment Factor of the top 1% of predictions.

The lower average precision value obtained from training with PubChem (0.16) compared to DUD-E (0.64) and DEKOIS 2.0 (0.62) is consistent with the expectation that highly imbalanced positive and negative sets lead to an underestimation of true positives. Specifically, the negative set is 1400 times as large as the positive set for PubChem whereas it is typically only 60 and 30 times as large for DUD-E and DEKOIS 2.0, respectively.

#### PrePCI Performance on an independent drug-target interaction gold standard data set

To compare PrePCI to current state-of-the-art matrix factorization methods, we performed 10-fold cross-validation with four benchmark datasets originally created by Yamanishi et al.<sup>107</sup> and recently updated by Liu et al.<sup>83</sup>, where each dataset contains PCI data for a separate class of protein targets: Enzymes, Nuclear Receptors, GPCRs and Ion channels (Table 2.S5 and Methods). For each class we calculated scores for all possible protein-compound pairs and obtained similar

performance as with PubChem-derived data (Table 2.S6). Despite quite high AUROC scores, PrePCI performance is less impressive than the other methods that utilize many more tunable parameters and don't rely on the availability of PDB template complexes (Table 2.S6). However, given the relative simplicity of PrePCI, its ability to provide interaction models for its predictions, and its proteome-wide applicability, its good performance within specific protein classes underscores its utility for more focused studies as illustrated in the following chapter.

### **2.3 The PrePCI database – PrePCI/DB**

PrePCI predictions are available through a web-hosted searchable database (PrePCI/DB) at <https://honiglab.c2b2.columbia.edu/prepci.html>. PrePCI/DB contains predictions for ~5 billion PCIs involving 6.8M compounds and 75,643 CDD domains representing 19,797 human proteins. Users can query the database for proteins (with UniProt Accession ID or gene name) or for compounds (PDB compound ID, PubChem CID or SMILES) to obtain PrePCI predictions for compounds and targets, respectively. Searching by protein will return a list of PDB compounds predicted to bind the protein by either structural similarity (Figure 2.1B), sequence similarity (Figure 2.1A), or both, along with the corresponding LT-scanner scores, BLAST e-values and PrePCI likelihood ratios (Figure 2.7). From our benchmarking results, we found that the LRs of 1,400,000, 15,000 and 190 correspond to FPR values of  $10^{-4}$ ,  $10^{-3}$ , and  $10^{-2}$  respectively. While predictions with higher LRs are more likely to be correct, predictions with lower LRs can still present a rich resource of plausible novel PCIs, particularly those with high LT-scanner scores which are more likely to contain evolutionarily conserved local structural similarity and thus constitute novel PCIs which may not be apparent from sequence analysis alone. The “Click to view PCI” icon triggers the website to display interactive JSMol windows for visualization of the

predicted binding interface as well the structural superposition of the query protein model and the PDB template complex. PDB-formatted files for both the interaction model and the structural superposition can be downloaded for further analysis including more detailed docking studies, as described in the following chapter. Additional similar compounds can be retrieved by clicking on the “Click to Find Other PCIs” icon which will open a new tab containing all PubChem compounds that are similar to the selected PDB compound. Together with the interaction visualization windows, this two-step procedure allows the user to evaluate predicted PCI interfaces before considering additional compounds likely to bind in a similar mode. Alternatively, users can query the database for a compound in the form of a PDB ID, PubChem CID or SMILES string. Case studies presented in the following chapter illustrate how both strategies – querying PrePCI/DB for predicted targets and for compounds – can be used to discover novel therapeutically interesting lead compounds and generate novel biological hypotheses.



**Figure 2.7: PrePCI webpage output.**

Protein query search: (A) The top of the webpage displays search criteria and the number of PCIs predicted. (B) Two JSmol windows display the query protein-compound interaction model (left) and query-template superposition (right) which a user may manipulate. (C) A table displays the PDB compounds predicted to interact with the query protein and their corresponding LT-scanner scores, BLAST e-values and PrePCI LRs. PCIs for chemically similar compounds: The “Click to View PCI” triggers the webpage to display the interaction and superposition models (panel B) while “Click to Find Similar Compounds” opens (D) a new webpage listing PubChem compounds similar to the query PDB compound, which can in turn be used to search for target proteins via the “Click to Find PCI” button.

## 2.4 Materials and Methods

### Template and Model selection

LT-scanner requires databases of query protein structure models and experimentally resolved holo-structures of protein-ligand co-complexes. To select a representative set of template PDB holo-structures, all PDB files in the PDB ligand expo (<http://ligand-expo.rcsb.org/>) were parsed to identify protein chains bound to ligands, and all chains were mapped to their respective UniProt IDs using the SIFTS database<sup>108, 109</sup>. Chains with more than one corresponding UniProt ID, commonly chimeric fusion proteins, as well as chains that did not map to a UniProt ID were excluded. X-ray crystal structures and cryo-EM structures with resolution less than 4Å and 4.5Å, respectively, were removed, and, when a PCI was represented more than once, the highest resolution complex was retained. This procedure yielded 55,994 unique template PCIs between 17,705 proteins and 25,613 compounds after removing compounds with molecular weight < 200 Da and fewer than 6 heavy atoms.

### Model Databases

An essential component of LT-scanner is a database of structural models for most query proteins and their constituent domains in the human proteome. To date, we have relied on our PrePMod database of homology models<sup>18</sup>. For the human reference proteome<sup>102, 103, 110</sup> (<https://www.uniprot.org/proteomes/UP000005640>), structural models for full-length sequences and protein domains as defined by the conserved domain database (CDD)<sup>111</sup> were constructed as follows.

### *PrePMod*

BLAST<sup>112</sup> was used to identify proteins in the PDB with sequences similar to the query sequence. For query-template sequence pairs with a BLAST e-value  $\leq 10^{-12}$ , a homology model for the query sequence was created with Nest<sup>113</sup>, our lab's in house algorithm which generates homology models from PDB template structures by iteratively mutating and refining residues in the template structure until the protein sequence is converted into the query protein. If no template was identified, remote sequence homologs within the PDB were identified by HHblits<sup>114</sup> with 5 iterations, and, if a template with an e-value  $\leq 10^{-12}$  was identified, a homology model was created with Nest<sup>113</sup>. Otherwise, a homology model for the query sequence was not created. This process yielded PrePMod, a protein model database containing 76,816 domain models for 17,150 human proteins.

### *AlphaFold/CDD (AF/CDD)*

Models for full length human query proteins were taken from the AlphaFold Protein Structure Database (AF)<sup>102, 103</sup>. Models for the constituent CDD domains for each of the proteins were excised from the full-length model sequence by retaining coordinates of atoms corresponding to residues defined as part of the domain by CDD. For proteins with more than 2700 residues, AF provides multiple sequence-redundant models by dividing the full-length protein sequence into overlapping 1400 residue long segments beginning every 200 residues. In such cases, multiple models may span the same domain. To identify a representative model for each domain in these cases, the pLDDT scores (the per-residue confidence metric) for each residue were summed across the CDD domains<sup>111</sup> of each model which fully contained the domain. The model with the largest total pLDDT was chosen as the representative for each domain. This procedure yielded 89,645 domain models for 20,526 proteins.

Altogether, the combination of PrePMod and AF/CDD provides 166,461 models for 90,308 domains for 20,599 proteins. This constitutes a significant (~15-20%) increase in structural coverage which now includes at least one representative for nearly every coding gene in the human proteome<sup>110</sup>.

### Protein and Chemical Similarity Methods

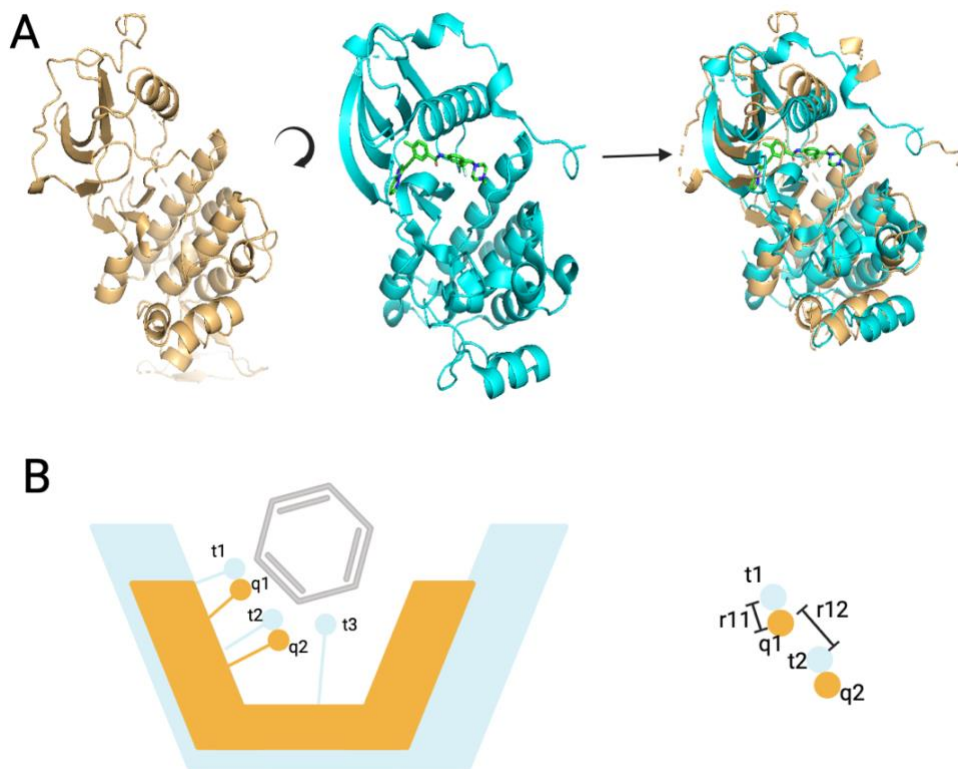
#### *LT-scanner*

The LT-scanner algorithm which uses structure alignment to relate query proteins to PDB template complexes has been described previously and is illustrated in Figure 2.8<sup>91</sup>. Briefly, the structure alignment program *ska*<sup>97</sup> identifies templates (T) from PDB protein-compound complexes for a query protein model (Q). T and Q are considered structurally similar when their protein structural distance (PSD)<sup>97</sup> is less than 0.8. The structural alignment performed by *ska* rotates Q into the coordinate frame of T, creating a putative interface between Q and the co-complexed compound (C). Hydrogen atoms are added to T and Q using the Open Babel Package with a pH of 7 and standard pK<sub>a</sub>s<sup>115</sup>. Four types of interactions between T and C, and Q and C, are identified according to the following definitions: 1) hydrogen bonds (distance < 3.5Å between heavy atoms and angle > 120), 2) aromatic-aromatic (distance ≤ 5Å), 3) ion pairs (distance ≤ 5Å), and 4) van der Waals contacts ( $0.5 \cdot r_{vdw} < \text{distance} < 1.2 \cdot r_{vdw}$ ) where  $r_{vdw}$  is the sum of the van der Waals radii for the two interacting atoms as defined in the Open Babel parameter set. The extent to which Q is able to recapitulate the intermolecular interactions formed between T and L is calculated by a similarity score, SIM(QC, TC) defined as:

$$S_{QC:TC} = \sum_u^{n_T} \sum_w^{n_Q} m_{uw} e^{-\gamma r_{uw}^2}$$

$$SIM(QC, TC) = \frac{S_{QC:TC}}{\max(S_{QC:QC}, S_{TC:TC})}$$

where  $u$  and  $w$  are indices of interfacial atoms of the template and query protein respectively,  $r_{uw}$  is the distance between atom  $u$  and atom  $w$ ,  $n_T$  and  $n_Q$  are the number of interfacial atoms in the template and query respectively,  $\gamma$  is an adjustable weighting parameter empirically chosen to be 0.7 and  $m_{uw}$  acts as a matching term which is 1 if atoms  $u$  and  $w$  are capable of forming a similar type of interaction (eg hydrogen bond donors) and 0 otherwise. Though distances for all pairwise combinations of template and query protein atoms are included in the  $S$  score, the use of the gaussian weighting term ensures that only pairs of atoms that are in geometrically similar positions significantly contribute to the  $S$  score, as the contribution of two distant atoms rapidly reduces to zero as interatomic distance increases. By normalizing  $S_{TQ}$  by  $\max(S_{TC:TC}, S_{QC:QC})$ , the resulting  $SIM$  score effectively counts the fraction of interactions made by the template that are recapitulated by the query while penalizing deviations in the position of analogous atoms (ie reducing  $S_{TC:QC}$  for atom pairs with  $r_{uw} > 0$  relative to the corresponding pair in  $S_{TC:TC}$  or  $S_{QC:QC}$ ) and without explicitly assigning template atoms to query atoms. For each potential protein-compound interaction, the LT-scanner score is defined as the maximal observed  $SIM$  score between the query protein and the compound. LT-scanner was applied to both the PrePMod and AF/CDD model databases. In cases where PrePMod and AF/CDD contain models for the same protein/domain, the query model that obtains the higher LT-scanner score was included in the LT-scanner evaluation analyses (Table 2.S2).



**Figure 2.8: Overview of LT-scanner Algorithm.**

(A) LT-scanner begins with a set of structural models, including homology models, AlphaFold models and experimental structures from the PDB, for each query protein in the human proteome (orange protein) as well as a set of protein-compound template holo-structures (blue protein with green ligand). Query protein models are superimposed on template holo-structures using *ska*, thereby rotating the query protein into the same coordinate frame as the template and creating a putative interface between the query protein and the template ligand. (B) Interfacial atoms in the template ( $t_i$ ) and the query ( $q_j$ ) are identified as described in Methods. All pairwise distances between template interfacial atoms and query interfacial atoms ( $r_{ij}$ ) are calculated and used as arguments for the LT-scanner SIM score described in Methods. Adapted from ref<sup>91</sup>.

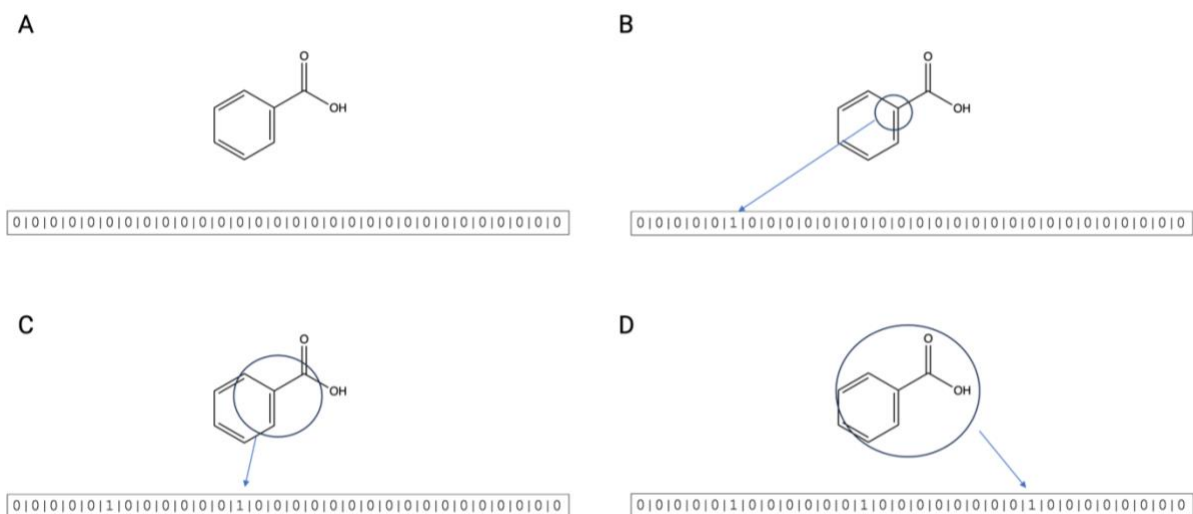
*Sequence similarity*

For each of the structures in the holo-structure template database, BLAST was run using the sequence of the PDB chain as a query against the human reference proteome (UP000005640)<sup>110</sup>. For a given query PCI, the PDB template complex containing the query compound which yielded with the smallest e-value relative to the query protein was identified. The

PCI was assigned an interaction sequence score of  $-\ln(\text{e-value})$ . E-values of 0 were re-assigned to  $1e-181$ , the smallest non-zero e-value obtained from the BLAST results, to prevent undefined scores. The sequence similarity component of PrePCI yields predictions for 17,864 proteins (Table S2.1).

### *Chemical similarity*

Chemical structure data for ~110M compounds and 26K PDB compounds was obtained from PubChem<sup>100, 101</sup> and the PDB<sup>94</sup> in SMILES format<sup>116</sup>. Rdkit<sup>117</sup> was used to compute 1024-bit Morgan2 fingerprints<sup>98</sup> for each compound (Figure 2.9), and Tanimoto coefficients<sup>99</sup> were computed for each PDB-PubChem compound pair. The reliability of inferring novel compounds from known compounds has been found to deteriorate below Tanimoto Coefficients of 0.5<sup>118</sup> so only those pairs of compounds with  $TC \geq 0.5$  were retained, yielding 6,835,528 Pubchem compounds similar to at least one PDB compound associated with a protein co-complex structure. Overall, PrePCI provides predictions for these 6.8M compounds.



### Figure 2.9: Overview of Morgan Chemical Fingerprint Algorithm.

(A) Beginning with a small molecule of interest, a fixed length vector where each vector position denotes the presence of a specific molecular fragment is initialized with all zeros. (B) An atom in the molecule is chosen (circle) and the presence of the atom is marked as a one in its corresponding vector position. (C) A larger fragment is identified by considering atoms adjacent to the starting atom (circle) and the presence of this fragment is similarly denoted as a one in its respective position in the fingerprint vector. (D) This procedure of expanding the topological radius of the fragment is continued for a finite number of iterations and repeated for each atom in the molecule, yielding a bitvector fingerprint which denotes the specific fragments which comprise the molecule. Adapted from ref <sup>98</sup>.

### Naïve Bayes Integration

A Naïve Bayes Classifier was trained to integrate scores into a single likelihood ratio (Figure 2.1). For each query PCI, a reference PDB compound, defined as the PDB compound which has the highest TC with the query compound of all PDB compounds predicted to bind to the query protein by either LT-scanner or sequence similarity, was identified. The chemical, structural, and sequence scores are then defined as 1) the TC between the query compound and the reference compound, 2) the LT-scanner score for the query protein-reference compound pair, and 3) the sequence score between the query protein and the most similar template protein from among complexes containing the reference compound, respectively. The number of bins was chosen as

10, 10, and 20 for chemical similarity, structural similarity, and sequence similarity respectively, with an additional NULL bin for each feature if no templates could be identified. The range of scores for each feature was divided into equal intervals based on the number of bins. Likelihood ratios for each feature and bin were computed as the ratio of the posterior odds and the prior odds that a given PCI is a true interaction:

$$LR = \frac{O_{posterior}}{O_{prior}} = \frac{P(TP|bin\ i)/P(TN|bin\ i)}{P(TP)/P(TN)}$$

where  $P(TP | bin\ i)$  and  $P(TN | bin\ i)$  are the probabilities that a given PCI is a true positive or a true negative respectively given they are yield a score which corresponds to bin  $i$  and  $P(TP)$  and  $P(TN)$  are the prior probabilities that a random protein-compound pair is a true positive PCI and true negative respectively. This relationship can be inverted to eliminate the priors using Bayes Theorem to yield the following:

$$LR = \frac{P(TP|bin\ i)/P(TN|bin\ i)}{P(TP)/P(TN)} = \frac{\frac{P(TP) \cdot P(bin\ i|TP)/P(bin\ i)}{P(TN) \cdot P(bin\ i|TN)/P(bin\ i)}}{\frac{P(TP)}{P(TN)}} = \frac{P(bin\ i | TP)}{P(bin\ i | TN)}$$

Which can be readily computed from training data as:

$$LR = \frac{P(bin\ i | TP)}{P(bin\ i | TN)} = \frac{\frac{N(bin\ i\ and\ TP)}{N(TP)}}{\frac{N(bin\ i\ and\ TN)}{N(TN)}} = \frac{N(bin\ i\ and\ TP) \cdot N(TN)}{N(bin\ i\ and\ TN) \cdot N(TP)}$$

The final likelihood ratio for a PCI defined as the product of the three component feature likelihood ratios:

$$LR(PCI) = \prod_{\substack{i=chemical, \\ structure, \\ sequence}} LR_i(bin(score))$$

## PubChem Benchmarking

Bioactivity data for each protein in the human proteome was downloaded as PubChem's "Tested Compounds" data from the "Chemicals and Bioactivities Data" section<sup>100, 101</sup>. The data was filtered to retain active, nonredundant experimental PCIs defined as "Active" for the "activity" feature or "<" or "=" for the "acqualifier" feature. This process yielded 1,122,699 PCIs involving 3,559 proteins and 642,498 compounds. Of the 642,498 PubChem compounds, 142,490 (22%) have Tanimoto Coefficient  $\geq 0.5$  with at least one PDB compound, and 2,926 of the 3,559 proteins have experimental evidence supporting an interaction with at least one of the 142,490 compounds. After filtering, the true positive set comprised 285,108 PCIs. The true negative set was defined as the remaining 416,640,632 possible PCIs (2926 proteins • 142,490 compounds – 285,108 true positives) not identified as true positives in Pubchem.

In the above benchmarking approach, all pairwise protein-compound pairs were considered in the training and test set, irrespective of whether PrePCI is effectively able to make a prediction for the specific protein-compound pair. Consequently, there were many PCIs for which PrePCI could not identify a reference compound and accordingly could not score the protein-compound pair effectively. To better evaluate the accuracy of PrePCI's predictions, we removed PCIs for which PrePCI does not make a prediction from the positive and negative sets which resulted in 204,919 true positive PCIS and 62,414,150 true negative protein-compound pairs which are 72% and 15% the size of the original PubChem benchmark set. Further, to evaluate PrePCI's performance using a more balanced dataset, we randomly sampled 2,049,190 of the 62,414,150 true negative interactions such that ratio of negatives to positives in each fold was 10:1.

In all three of the above cases, we split both the positive and negative sets into 10 mutually disjoint subsets and, using each subset in turn as a test set, trained LRs using the PCIs from the

remaining 9 sets as a training set. PCIs in the test set were scored and ranked by their composite LR, and the AUROC and average precision were computed using scikit-learn.

### Benchmarking on an independent docking Dataset

Two widely used benchmarking datasets, the Directory of Useful Decoys Enhanced (DUD-E)<sup>104</sup> and the Demanding Evaluation Kits for Objective In Silico Screening 2.0 (DEKOIS 2.0)<sup>105</sup>, were used to further evaluate PrePCI and compare its performance to that of other existing methods. DUD-E and DEKOIS 2.0 contain PCIs for both active compounds and property matched decoy compounds. DUD-E contains 95 human proteins and over a million PCIs with 19K active compounds and 1.2M decoy compounds while DEKOIS 2.0 contains 69 human proteins and 87K PCIs with 2.6K active compounds and 77K decoy compounds. PCI active and decoy compounds for the 102 proteins in the DUD-E database were downloaded from <http://dude.docking.org/db/subsets/all/all.tar.gz><sup>104</sup>. PCI data for the proteins in the DEKOIS 2.0 database was downloaded from <http://www.pharmchem.uni-tuebingen.de/dekois/><sup>105</sup>. Because these datasets contain compounds whose maximal TC with a PDB compound is less than 0.5, 10 additional chemical similarity bins were added to extend the range of TCs down to 0 so that predictions could be made for all PCIs in the datasets.

Naïve Bayes classifiers were trained using leave-one-out cross-validation for the proteins in the DUD-E and DEKOIS 2.0 datasets where likelihood ratios were computed using active and decoy protein-compound pairs from all but one protein and evaluated on the active and decoy compounds for the protein withheld from training. PCIs with experimentally resolved structures in the PDB were included in training but excluded from test sets to remove recall bias. For each of the cross-validation folds, predictions for PCIs were ranked and ROC and precision-recall curves

were created using scikit-learn and matplotlib, and AUROC and average precision were computed using scikit-learn. In addition, the enrichment factor of the top 1% of predictions ( $EF_{0.01}$ ), a metric commonly used to estimate the degree to which true binders are enriched in the top-scoring predictions relative to random, was calculated as follows:

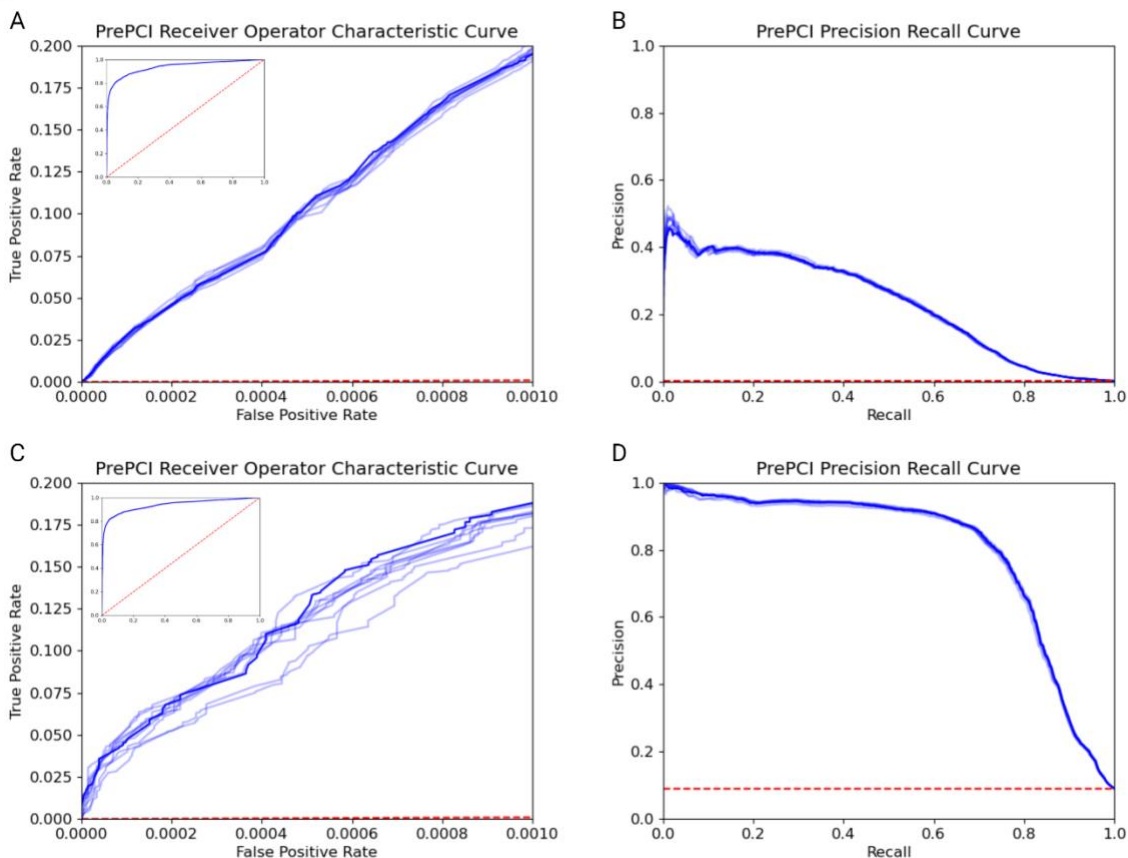
$$EF_{0.01} = \frac{N_{top}}{0.01 \cdot N_{active}}$$

where  $N_{top}$  is the number of true positives recovered in the top 1% of predictions and  $N_{active}$  is the total number of true positives in the test set.

#### Benchmarking on an independent drug target interaction gold standard dataset

To facilitate the comparative benchmarking of different protein-compound interactions within families, Yamanishi et al released a collection of 4 protein-compound interaction datasets, where each dataset contains experimental protein-compound interaction data for one of four protein classes, enzymes, ion channels, nuclear receptors and G-protein coupled receptors<sup>107</sup>. Updated versions of the protein class datasets compiled by Yamanishi et al.<sup>107</sup> and updated in Liu et. al.<sup>83</sup> were obtained from [https://github.com/intelligence-csd-auth-gr/DTI\\_MDMF2A/tree/main/datasets\\_my](https://github.com/intelligence-csd-auth-gr/DTI_MDMF2A/tree/main/datasets_my). KEGG Compound IDs were mapped to SMILES strings using the Pubchem Chemical Identifier Exchange Service (<https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi>). KEGG protein IDs were mapped to Uniprot IDs using the Uniprot ID mapping tool (<https://www.uniprot.org/id-mapping>). PCIs present in each dataset were considered true positives while the remaining all-on-all protein-compound pairs were considered true negatives. We performed 10-fold cross-validation for each protein class and calculated performance statistics.

## 2.5 Supplemental Information



**Figure 2.S1: Performance of PrePCI predictions on unbalanced and balanced PubChem protein-compound interaction experimental data.**

We evaluated the PrePCI on true positive and true negative sets for which PrePCI makes is able to make predictions. This restriction reduced the size of the true positive set from 285,108 to 204,919 PCIs and reduced the size of the true negative set from 416,640,632 to 62,414,150. A) Receiver operating characteristic curve and (B) Precision Recall curve for each of the 10 cross-validation folds using the restricted set of PCIs were obtained. Curves for the median area under the ROC curve (AUROC, A) and average precision (B) across all cross-validation folds are darker, while curves for all cross-validation folds (lighter blue) are included to display the range of results obtained from the individual folds. PrePCI's average AUROC and Average Precision on the PubChem dataset are  $0.936 \pm 0.001$  and  $0.233 \pm 0.002$  respectively. Figure S1A,B suggest that when PrePCI is able to identify a template PCI, it readily prioritizes positive interactions over negative protein-compound pairs thus adding confidence in the value of the database.

Finally, because the assumption that all protein-compound pairs without bioactivity data are true negatives contradicts the application of identifying novel PCIs, we further subsampled the negative set of scored PrePCI predictions so that the ratio of negatives to positives in each fold was 10:1 to reflect the expectation that true positive PCIs are relatively uncommon and recalculated (C) ROC and (D) Precision-Recall curves for this more balanced dataset. The resulting ROC curves (Figure 2.S1C) are highly similar to those obtained by considering all PrePCI

predictions (Figure 2.S1A) with a mean AUROC of  $0.935\pm 0.001$ . In contrast, the PR curves dramatically improved, with mean average precision of  $0.798\pm 0.003$ . This reflects both PrePCI's ability to prioritize true positive interactions over a wide range of protein targets, as well as the inflationary effect that restricting class imbalance by subsampling has on both the AUPRC and random precision. For practical purposes, we consider the results obtained by considering only PCIs which PrePCI can effectively score (Figure 2.S1A,B) as most representative of the predictions within the PrePCI/DB database.

LT-scanner score threshold (LR)	Sequence Threshold (LR)	LT-scanner PCI Predictions	Sequence PCI Predictions	PCI predictions unique to LT-scanner	PCIs predicted by BOTH LT-scanner and Sequence	PCI predictions unique to Sequence	PCIs predicted by Union of LT-scanner and Sequence
0.6 (141)	186.2791 (135)	101,121	62,153	72,173	28,948	33,205	134,326
0.5 (42)	123.4186 (26)	258,498	115,905	197,085	61,413	54,492	312,990
0.4 (10)	60.55799 (10)	675,105	423,786	446,875	228,230	195,556	870,661
0.3 (3.8)	39.60446 (4.0)	1,566,204	749,001	1,040,624	525,580	223,421	1,789,625
0.2 (2.1)	-2.30259 (2.2)	4,449,502	2,385,839	2,652,348	1,797,154	588,685	5,038,187

**Table 2.S1: Comparison of the number of PCIs predicted by LT-scanner and sequence-based metrics at comparable LRs**

LT-scanner thresholds (Column 1) and Sequence similarity score thresholds (column 2) corresponding to similar LRs were identified. Columns 3 and 4 indicate the number of PCI predictions with LT-scanner scores and sequence similarity scores exceeding their respective threshold. Column 5 indicates the number of PCIs predicted by LT-scanner but not sequence similarity, and conversely, column 7 indicates the number of PCIs predicted by sequence similarity but not LT-scanner. Column 6 indicates the number of PCIs predicted with LT-scanner and sequence similarity scores greater than their respective thresholds (i.e. the intersection of sequence and LT-scanner predictions) while Column 8 indicates the number of PCIs predicted by either LT-scanner or sequence similarity (i.e. the union of the LT-scanner and sequence similarity predictions).

LT-scanner threshold	PCIs predicted by PrePMod	PCIs predicted by AF/CDD	PCI Predictions Unique to PrePMod	PCIs predicted by both PrePMod and AF/CDD	PCI Predictions Unique to AF/CDD	Union of LT-Scanner predictions with PrePMod and AF/CDD
0.6	63,730	76,765	24,356	39,374	37,391	101,121
0.5	153,660	210,225	48,273	105,387	104,838	258,498
0.4	401,728	561,365	113,740	287,988	273,377	675,105
0.3	1,044,492	1,313,141	253,063	791,429	521,712	1,566,204
0.2	3,043,045	3,410,595	1,038,907	2,004,138	1,406,457	4,449,502

**Table 2.S2: Comparison of the number of PCIs predicted by LT-scanner using PrePMod and AF/CDD model databases.**

At different LT-Scanner LR thresholds (Column 1), the number of PCI predictions obtained using the model databases indicated are provided. Columns 2 and 3 indicate the number of PCIs predicted by LT-scanner using PrePMod and AF/CDD, respectively. Column 4 indicates the number of PCIs predicted when using PrePMod models that are not found with AF/CDD models and conversely, column 6 indicates the number of PCIs detected by AF/CDD and not PrePMod. Column 5 indicates the number of PCIs identified by both PrePMod and AF/CDD models (i.e. the intersection of PrePMod and AF/CDD predictions) while Column 5 indicates the total number of PCIs predicted by LT-scanner using both model sets (i.e. the union of PrePMod and AF/CDD predictions).

	<i>FRAGSITE</i>		<i>FINDSITE_comb_2.0</i>		<i>AutoDock Vina</i>		<i>PrePCI</i>	Average Precision
	EF 0.01	AUPR	EF 0.01	AUPR	EF 0.01	AUPR	EF 0.01	
<i>aa2ar</i>	0.21	0.02	0	0.01	1.87	0.03	<b>58.75</b>	<b>0.7</b>
<i>abl1</i>	51.1	0.73	47.92	0.62	3.31	0.03	<b>58.52</b>	<b>0.75</b>
<i>ace</i>	44.68	0.68	49.41	0.71	2.13	0.02	<b>57.86</b>	<b>0.78</b>
<i>aces</i>	10.38	0.09	4.63	0.05	7.28	0.07	<b>51.77</b>	<b>0.64</b>
<i>ada</i>	32.26	0.4	41.17	0.54	0	0.02	<b>59.14</b>	<b>0.86</b>
<i>ada17</i>	62.03	<b>0.8</b>	38.58	0.48	2.82	0.02	<b>62.45</b>	0.77
<i>adrb1</i>	6.07	0.07	3.24	0.03	4.45	0.04	<b>55.47</b>	<b>0.59</b>
<i>adrb2</i>	15.58	0.11	6.5	0.05	0.43	0.02	<b>45.22</b>	<b>0.47</b>
<i>akt1</i>	47.1	0.74	38.95	0.49	8.2	0.05	<b>57.44</b>	<b>0.86</b>
<i>akt2</i>	50.43	<b>0.76</b>	43.65	0.52	18.83	0.11	<b>53.98</b>	0.65
<i>aldr</i>	0.63	0.02	0.63	0.01	2.5	0.02	<b>53.38</b>	<b>0.61</b>
<i>ampc</i>	<b>4.17</b>	<b>0.04</b>	0	<b>0.04</b>	0	0.01 (-)	(-)	(-)
<i>andr</i>	10.78	0.09	14.14	0.11	0	0.01	<b>35.34</b>	<b>0.44</b>
<i>aofb</i>	0	0.02	0.82	0.01	1.64	0.04	<b>18.18</b>	<b>0.19</b>
<i>bace1</i>	50.53	0.62	48.67	0.58	4.23	0.04	<b>68.8</b>	<b>0.85</b>
<i>braf</i>	48.03	<b>0.69</b>	44.05	0.55	5.92	0.05	<b>51.02</b>	0.6
<i>cah2</i>	0.41	0.02	0	0.01	2.24	0.02	<b>53.45</b>	<b>0.65</b>
<i>casp3</i>	34.17	0.45	<b>44.18</b>	<b>0.66</b>	1	0.03	39.9	0.57
<i>cdk2</i>	25.11	0.3	18.36	0.15	2.32	0.03	<b>56.57</b>	<b>0.64</b>
<i>comt</i>	12.2	0.12	26.75	0.2	0	0.01	<b>60</b>	<b>0.68</b>
<i>cp2c9</i>	<b>42.5</b>	<b>0.5</b>	26.55	0.3	4.98	0.06	14.17	0.13
<i>cp3a4</i>	<b>32.94</b>	<b>0.34</b>	21.7	0.24	2.35	0.02	10.65	0.09
<i>csf1r</i>	39.76	0.37	33.76	0.36	3.01	0.02	<b>65.24</b>	<b>0.75</b>
<i>cxcr4</i>	<b>42.5</b>	0.27	0	0.03	0	0.01	37.5	<b>0.39</b>
<i>def</i>	<b>49.02</b>	<b>0.86</b>	24.5	0.33	0	0.01 (-)	(-)	(-)
<i>dhi1</i>	0	0.01	0	0.02	0.91	0.02	<b>48.32</b>	<b>0.53</b>
<i>dpp4</i>	4.5	0.03	8.08	0.05	0.38	0.01	<b>67.82</b>	<b>0.72</b>
<i>drd3</i>	11.25	0.09	8.96	0.07	3.33	0.03	<b>34.38</b>	<b>0.33</b>
<i>dyr</i>	23.81	0.13	26.84	0.27	3.9	0.03	<b>66.67</b>	<b>0.81</b>
<i>egfr</i>	<b>59.04</b>	<b>0.74</b>	54.74	0.66	1.29	0.02	55.19	0.69
<i>esr1</i>	21.41	0.3	12.04	0.1	1.31	0.03	<b>56.6</b>	<b>0.87</b>
<i>esr2</i>	18.8	0.28	12.29	0.12	0.55	0.01	<b>56.35</b>	<b>0.89</b>
<i>fa10</i>	13.41	0.18	11.34	0.15	5.39	0.08	<b>55.43</b>	<b>0.89</b>
<i>fa7</i>	32.46	0.35	6.97	0.1	9.59	0.09	<b>53.57</b>	<b>0.79</b>
<i>fabp4</i>	0	0.03	2.12	0.05	0	0.04	<b>61.9</b>	<b>0.84</b>
<i>fak1</i>	50	0.9	44.59	0.69	0.99	0.02	<b>55.67</b>	<b>0.95</b>
<i>fgfr1</i>	23.74	<b>0.69</b>	28.34	0.67	8.72	0.16	<b>54.74</b>	<b>0.69</b>
<i>fkbl1a</i>	4.5	0.03	0	0.05	1.81	0.03	<b>53.15</b>	<b>0.9</b>
<i>fnta</i>	<b>31.93</b>	<b>0.33</b>	31.6	0.32	0.51	0.01	24.49	0.22
<i>fpps</i>	56.47	0.6	55.34	0.54	0	0.01	<b>65.82</b>	<b>0.69</b>

<i>gcr</i>	29.07	0.34	34.21	0.34	0	0.01		<b>41.8</b>	<b>0.45</b>
<i>glcm</i>	18.52	0.14	3.66	0.09	0	0.01		<b>28.3</b>	<b>0.27</b>
<i>gria2</i>	34.81	0.35	<b>44.9</b>	<b>0.43</b>	3.79	0.02		29.3	0.27
<i>grik1</i>	35.64	0.39	40.89	0.47	1	0.02		<b>53.47</b>	<b>0.65</b>
<i>hdac2</i>	28.11	0.41	11.87	0.13	1.08	0.04		<b>49.73</b>	<b>0.67</b>
<i>hdac8</i>	26.47	<b>0.41</b>	15.32	0.15	4.71	0.05		<b>41.92</b>	<b>0.41</b>
<i>hivint</i>	0	<b>0.03</b>	0	0.01	<b>1.01</b>	0.02	(-)		(-)
<i>hivpr</i>	<b>48.69</b>	<b>0.61</b>	34.72	0.42	0.37	0.02	(-)		(-)
<i>hivrt</i>	<b>2.37</b>	<b>0.05</b>	1.18	0.02	0.89	0.02	(-)		(-)
<i>hmdh</i>	12.94	0.18	11.78	0.2	0	0.02		<b>53.61</b>	<b>0.72</b>
<i>hs90a</i>	12.5	0.12	3.43	0.08	0	0.01		<b>64.38</b>	<b>0.81</b>
<i>hvk4</i>	4.35	0.08	0	0.02	0	0.02		<b>52.81</b>	<b>0.9</b>
<i>igf1r</i>	60.81	<b>0.8</b>	46.14	0.6	2.04	0.03		<b>62.33</b>	<b>0.8</b>
<i>inha</i>	<b>41.86</b>	<b>0.53</b>	40.27	0.47	9.48	0.05	(-)		(-)
<i>ital</i>	40.58	0.5	25.44	0.21	0	0.02		<b>60.45</b>	<b>0.62</b>
<i>jak2</i>	<b>58.88</b>	<b>0.86</b>	34.59	0.46	2.8	0.03		54.46	0.63
<i>kif11</i>	0	0.02	0.86	0.03	0	0.01		<b>64.49</b>	<b>0.9</b>
<i>kit</i>	43.37	0.6	50.06	0.6	1.81	0.02		<b>56.71</b>	<b>0.7</b>
<i>kith</i>	50.88	<b>0.99</b>	<b>51</b>	0.94	21.1	0.19		45.61	0.54
<i>kpcb</i>	29.63	0.37	<b>48.3</b>	<b>0.54</b>	17.09	0.17		39.26	0.47
<i>lck</i>	54.29	0.68	47.85	0.58	6.9	0.05		<b>54.81</b>	<b>0.74</b>
<i>lkha4</i>	13.45	0.16	8.2	0.11	1.17	0.04		<b>57.83</b>	<b>0.89</b>
<i>mapk2</i>	36.63	0.42	27.94	0.33	6.98	0.06		<b>62.24</b>	<b>0.77</b>
<i>mcr</i>	27.66	0.3	<b>31.09</b>	<b>0.36</b>	0	0.01		20	0.27
<i>met</i>	54.22	0.67	39.78	0.47	8.44	0.08		<b>69.75</b>	<b>0.9</b>
<i>mk01</i>	54.43	0.83	54.75	<b>0.84</b>	0	0.03		<b>61.64</b>	0.79
<i>mk10</i>	<b>47.12</b>	<b>0.61</b>	41.36	0.51	2.89	0.02		42.27	0.49
<i>mk14</i>	45.16	0.57	35.11	0.44	4.32	0.04		<b>58.02</b>	<b>0.68</b>
<i>mmp13</i>	62.94	0.75	40.73	0.5	5.07	0.03		<b>63.96</b>	<b>0.79</b>
<i>mp2k1</i>	41.32	0.44	23.87	0.18	0	0.01		<b>53.72</b>	<b>0.61</b>
<i>nos1</i>	1	0.02	0	0.01	0	0.02		<b>41</b>	<b>0.36</b>
<i>nram</i>	0	0.01	<b>1.02</b>	<b>0.08</b>	0	0.01	(-)		(-)
<i>pa2ga</i>	40.4	0.54	33.62	0.41	1.02	0.02		<b>50</b>	<b>0.65</b>
<i>parp1</i>	34.45	0.44	26.02	0.33	14.98	0.12		<b>56.97</b>	<b>0.66</b>
<i>pde5a</i>	61.06	0.8	<b>68.39</b>	<b>0.85</b>	0.75	0.01		46.97	0.49
<i>pgh1</i>	4.1	0.04	2.05	0.02	1.02	0.02		<b>30.77</b>	<b>0.28</b>
<i>pgh2</i>	6.44	0.04	5.28	0.04	3.44	0.03		<b>41.01</b>	<b>0.4</b>
<i>plk1</i>	60.75	<b>0.78</b>	55.17	0.64	0	0.02		<b>61.9</b>	0.72
<i>pnph</i>	62.14	<b>0.77</b>	57.87	0.73	0.96	0.02		<b>65.05</b>	0.7
<i>ppara</i>	9.92	0.14	1.34	0.03	0.81	0.02		<b>50.54</b>	<b>0.74</b>
<i>ppard</i>	17.08	0.17	7.06	0.08	2.08	0.04		<b>47.03</b>	<b>0.63</b>
<i>pparg</i>	12.4	0.15	12	0.1	2.69	0.05		<b>46.81</b>	<b>0.58</b>
<i>prgr</i>	10.58	0.11	18.13	0.14	1.03	0.03		<b>33.68</b>	<b>0.41</b>

<i>ptn1</i>	14.62	0.15	2.3	0.07	6.9	0.05	<b>47.97</b>	<b>0.54</b>
<i>pur2</i>	38	0.64	<b>54.88</b>	<b>0.96</b>	2.03	0.08	51.06	0.62
<i>pygm</i>	14.29	0.09	0	0.06	10.43	0.06	<b>37.66</b>	<b>0.56</b>
<i>pyrd</i>	1.8	0.03	0	0.03	0	0.03	<b>62.5</b>	<b>0.83</b>
<i>reni</i>	56.73	0.65	42.06	0.52	0.96	0.02	<b>66.99</b>	<b>0.73</b>
<i>rock1</i>	<b>49</b>	<b>0.64</b>	44.98	0.53	1	0.03	38	0.46
<i>rxra</i>	<b>51.15</b>	<b>0.72</b>	31.15	0.42	0.76	0.02	48.82	0.64
<i>sahh</i>	55.56	<b>0.98</b>	<b>55.76</b>	0.96	0	0.01	52.38	0.79
<i>src</i>	55.92	0.69	46.54	0.59	1.14	0.02	<b>63.01</b>	<b>0.78</b>
<i>tgfr1</i>	60.9	<b>0.84</b>	54.33	0.68	5.28	0.03	<b>64.57</b>	0.77
<i>thb</i>	4.85	0.05	0	0.02	0	0.02	<b>66</b>	<b>0.69</b>
<i>thrb</i>	54.88	0.66	50.56	0.65	0.87	0.02	<b>55.04</b>	<b>0.81</b>
<i>try1</i>	9.8	0.09	10.01	0.09	2.89	0.04	<b>46.55</b>	<b>0.61</b>
<i>trybl</i>	51.35	<b>0.91</b>	<b>51.97</b>	0.89	12.15	0.14	27.7	0.3
<i>tysy</i>	14.68	0.29	10.16	0.11	2.77	0.04	<b>53.7</b>	<b>0.77</b>
<i>urok</i>	<b>54.32</b>	<b>0.7</b>	27.79	0.31	0.62	0.02	45.86	0.62
<i>vgfr2</i>	48.9	<b>0.68</b>	40.15	0.49	5.39	0.05	<b>49.37</b>	0.61
<i>wee1</i>	43.14	0.65	31.12	0.43	61.27	<b>1</b>	<b>61.39</b>	<b>1</b>
<i>xiap</i>	0	0.04	0	0.02	0	0.01	<b>53.61</b>	<b>0.92</b>

**Table 2.S3: AUPRC and EF performance of PrePCI, FRAGSITE, FINDSITEcomb2.0 and AutodockVina on the DUDE dataset.**

Bold values indicate the highest AUPRC or EF of the four methods for the corresponding protein. (-) indicates that the corresponding PDB file contains a non-human protein and we did not attempt to make predictions using PrePCI. <sup>a</sup>Values taken from Table S2 of reference <sup>93</sup>.

<i>PDB</i>	<i>vScreenML<sup>a</sup></i>	<i>FINDSITE comb 2.0<sup>l</sup></i>	<i>FRAGSITE<sup>a</sup></i>	<i>PrePCI</i>
<i>3hng</i>	10.3	25.8	20	<b>27.50</b>
<i>1hov</i>	2.5	<b>31</b>	30	25.00
<i>3ny9</i>	0	0	<b>22.5</b>	20.00
<i>3kk6</i>	10.8	5.2	<b>20</b>	(-)
<i>1nhz</i>	5.4	<b>28.4</b>	27.5	15.00
<i>1xp0</i>	8.1	<b>31</b>	22.5	28.21
<i>1z11</i>	0	<b>10.3</b>	7.5	0.00
<i>3tfq</i>	8.6	0	0	<b>30.77</b>
<i>2oo8</i>	5.1	<b>28.4</b>	25	25.00
<i>1b8o</i>	7.50	<b>28.4</b>	27.5	(-)
<i>2w3l</i>	5.5	15.5	10	<b>30.00</b>
<i>1hw8</i>	24.6	5.2	<b>30</b>	<b>30.00</b>
<i>2afx</i>	0	0	0	<b>28.21</b>
<i>3ewj</i>	2.7	<b>31</b>	30	28.21
<i>3v8s</i>	18	15.5	<b>22.5</b>	<b>22.50</b>
<i>3eml</i>	7.7	0	10	<b>25.64</b>
<i>2z94</i>	0	0	0	(-)
<i>1uze</i>	21.4	10.3	5	<b>25.00</b>
<i>3klm</i>	5.4	5.2	5	<b>17.50</b>
<i>2wcg</i>	2.6	<b>15.5</b>	12.5	7.50
<i>1w4r</i>	0	20.7	0	<b>30.00</b>
<i>1r4l</i>	8.1	12.9	<b>22.5</b>	12.50
<i>1uou</i>	0	0	0	<b>27.50</b>

**Table 2.S4: EF performance of PrePCI, FRAGSITE, FINDSITEcomb2.0 and vScreenML on the DEKOIS dataset.**

Bold values indicate the highest EF of the four methods for the corresponding protein. (-) indicates that the corresponding PDB file contains a non-human protein and we did not attempt to make predictions using PrePCI. <sup>a</sup>Values taken from Table 5 of reference <sup>93</sup>.

**SI Table 5**

<i>Dataset</i>	<i>Number of compounds</i>	<i>Number of Proteins</i>	<i>Number of Interactions</i>
<i>Nuclear Receptor</i>	54	26	166
<i>G-Protein Coupled Receptor</i>	223	95	1096
<i>Ion Channel</i>	210	204	2331
<i>Enzyme</i>	445	664	4256

**Table 2.S5: Composition of the independent drug-target interaction gold standard set<sup>107</sup>**

SI Table 6

<i>Method</i>	<i>Dataset</i>			
	Nuclear Receptors	G-Protein Coupled Receptors	Ion Channels	Enzymes
<i>MSCMF</i>	0.882	0.962	0.982	0.961
<i>NRMFL</i>	0.882	0.972	<b>0.989</b>	0.981
<i>GRGMF</i>	0.891	<b>0.978</b>	0.988	0.982
<i>MF2A</i>	0.884	<b>0.978</b>	<b>0.989</b>	0.983
<i>DRLSM</i>	0.879	0.971	0.981	0.964
<i>DTINet</i>	0.797	0.916	0.938	0.839
<i>NEDTP</i>	0.846	0.953	0.981	0.97
<i>NetMD</i>	0.818	0.96	0.985	0.966
<i>Multi2Vec</i>	0.788	0.93	0.97	0.944
<i>MDMF2A</i>	<b>0.892</b>	<b>0.978</b>	<b>0.989</b>	<b>0.984</b>
<i>PrePCI*</i>	0.83	0.91	0.764	0.886

**Table 2.S6: Comparison of PrePCI performance to matrix factorization methods on the independent drug-target interaction datasets.**

AUROC results for the methods MSCMF, NRMFL, GRGMF, MF2A, DRLSM, DTINet, NEDTP, NetMD, Multi2Vec and MDMF2A taken from section S1 of Table 3 of Liu et. al.<sup>83</sup> \*PrePCI AUROCs obtained by considering all protein-compound pairs for which PrePCI makes a prediction. As a result, we obtain PrePCI scores for 135 (of 166) TPs and 917 (of 1,238) TNs on the nuclear receptor dataset, 138 (of 1,096) TPs and 2376 (of 20,089) TNs on the G-protein coupled receptor dataset, 605 (of 2,331) TPs and 4341 (of 40,509) TNs on the ion channel dataset, 1328 (of 4,256) TPs and 41,635 (of 291,224) TNs on the enzyme dataset. This reflects a current technical limitation of PrePCI that, in restricting our focus to compounds with TC  $\geq 0.5$  with a template PDB compound, we are unable to make predictions for compounds with only non-structural evidence for an interaction. This limitation complicates direct comparison to matrix-factorization approaches as PrePCI cannot make meaningful predictions for most PCIs in the dataset, however our results suggest that for PCIs where PrePCI is able to make a prediction, it is of similar confidence to existing methods.

## Chapter 3: Applications of PrePCI

### 3.1 Introduction

Having demonstrated PrePCI's performance on retrospective benchmarking sets and compared it to other similar methods, we subsequently sought to apply PrePCI prospectively to common biological and pharmacologic tasks. To illustrate potential applications of the PrePCI database, we present a number of case studies where PrePCI is used to suggest novel lead compounds for cancer targets, infer a possible direct target underlying methotrexate's mechanism of action and annotate protein function based on predicted interactions with cellular signaling molecules. Additionally, we present preliminary results integrating PrePCI and PrePPI to identify compounds predicted to bind at the interface of cancer master regulator binding proteins<sup>16, 23, 119</sup>. In all of the above cases, we focused particularly on PCIs with high structural scores but low sequence similarity scores to emphasize the contribution of structure in proteome-scale PCI prediction. However, although PrePCI is able to rapidly generate and score protein-compound interaction models, its use of rigid-body structural superposition to align pre-built homology models to template holo-structures can often yield interaction models which contain physically implausible features, such as clashing atoms, unpaired charges buried in hydrophobic patches and slightly misaligned binding sites. Thus, while the binding site of the derived model may share general structural similarity, an individual model may contain features which, if scored directly by docking or machine-learned scoring functions, may lead to erroneous predictions. For this reason, we believe PrePCI is most effectively applied as a hypothesis generation tool for identifying plausible structural relationships between proteins which can subsequently be evaluated more

rigorously using detailed computational and experimental techniques. As described below, we used a selection of physics-based methods from the Schrodinger suite of tools to more rigorously analyze several interesting PrePCI predictions.

### **3.1.1 Additional tools for PCI docking and affinity estimation**

#### Rigid-body Docking

Glide is a rigid-body docking program which iteratively samples numerous binding poses of flexible ligands within a fixed grid describing the protein receptor's pharmacophoric environment<sup>56, 57</sup>. Because it treats the protein as rigid and does not sample the configurational space of the protein, Glide is very fast, capable of generating and scoring binding poses within seconds and making it an ideal method for screening large libraries of compounds against high-confidence receptor geometries. The affinities predicted by Glide can be used to prioritize ligands for further analysis while the predicted binding mode can serve as an initial model for refinement. However, because rigid-body docking methods like Glide do not account for subtle structural changes that may occur upon ligand binding, inaccuracies in the receptor structure used for docking may lead to inaccuracies in both the predicted binding pose and affinity. We sought to mitigate such inaccuracies by preparing each interaction model in the presence of the ligand and performing a restrained minimization of the protein to relieve clashes prior to screening. However, for the most interesting PCIs, we used the following methods to sample protein configurational space more rigorously.

## Induced Fit Docking with Metadynamics (IFD-MD)

As mentioned above, protein side chains, and even backbones, can undergo significant reorientation relative to the unbound protein apo-structure upon ligand binding<sup>58</sup>. This reorientation can dramatically affect the accuracy of rigid-body docking methods which generally perform significantly better when using a ligand-bound holo-structure as the target receptor compared to using an apo-structure of the same protein<sup>58, 62</sup>. While the risk of using inaccurate query models in LT-scanner is likely mitigated through the use of multiple models for each protein, including homology models which can be based on holo-structure templates, it is still likely that most of the generated interaction models include some extent of steric clashing which, though not explicitly penalized in the LT-scanner scoring function, could interfere with more detailed biophysical simulations where ligand atoms might clash with the protein.

To address these induced fit effects, we used Schrodinger's Induced Fit Docking with Metadynamics (IFD-MD), a method which efficiently combines pharmacophore docking, molecular dynamics and energy-guided protein structure modeling to generate alternative protein conformations that more accurately capture the receptor geometry and binding pose for a given PCI<sup>58</sup>. IFD-MD has been rigorously benchmarked and found to identify the correct binding mode within its top 2 docked poses in 95% of cases<sup>58</sup>. Moreover, Free Energy Perturbation (FEP, see next section) studies based on IFD-MD derived initial poses have been found to accurately recapitulate ligand binding free energies in homology models when initially prepared with IFD-MD<sup>120</sup>. While IFD-MD is able to predict the correct binding pose for a query PCI, its estimation of binding affinities is limited by its approximate treatment of entropic effects. To more accurately estimate binding free energies for the most interesting PCIs, we turned to the following method, free energy perturbation.

## Free Energy Perturbation (FEP)

Accurate estimation of binding free energies is one of the central goals of computational chemistry. While direct estimation of a protein-ligand binding free energy can, in principle be obtained through a statistical analysis of unbiased molecular dynamics calculations, the simulation times required to sufficiently sample the configurational space render such a direct approach intractable<sup>121</sup>. Enhanced sampling techniques which bias the simulation towards unsampled regions of configuration space can be used to sample energetically unfavorable phase spaces so as to map the full energy landscape<sup>121-123</sup>. However, in addition to large computational costs, these methods typically require significant effort to choose the effective biasing potential and reaction coordinates that are suitable to the studied system<sup>121</sup>.

In contrast, Free Energy Perturbation (FEP) casts the problem of estimating binding free energies as a thermodynamic cycle in which an initial reference ligand, A is converted to a final ligand, B (Figure 3.1). As mentioned above, binding free energies between a protein and ligand A ( $\Delta G^A$ ) as well as between the protein and ligand B ( $\Delta G^B$ ) can in principle be estimated using long MD or metadynamic simulations. The difference between in binding free energies ( $\Delta\Delta G^{AB}$ ), between  $\Delta G^A$  and  $\Delta G^B$  then is simply  $\Delta G^B - \Delta G^A$ . However, it is more computationally efficient to simulate the alchemical conversion of ligand A into B both in solvent ( $\Delta G^{A \rightarrow B}_{\text{solvent}}$ ) and when bound to the protein ( $\Delta G^{A \rightarrow B}_{\text{complex}}$ ). Because free energy is a state function, the total free energy change around the thermodynamic cycle is zero and the sum of the free energy changes along each leg of the cycle can be rearranged to yield the change in binding free energy on converting ligand A to ligand B as follows:

$$0 = \Delta G_{Bind}^A - \Delta G_{Bind}^B + \Delta G_{Complex}^{A \rightarrow B} - \Delta G_{Solvent}^{A \rightarrow B}, \quad \therefore$$

$$\Delta \Delta G_{Bind}^{A \rightarrow B} = \Delta G_{Bind}^B - \Delta G_{Bind}^A = \Delta G_{Complex}^{A \rightarrow B} - \Delta G_{Solvent}^{A \rightarrow B}$$

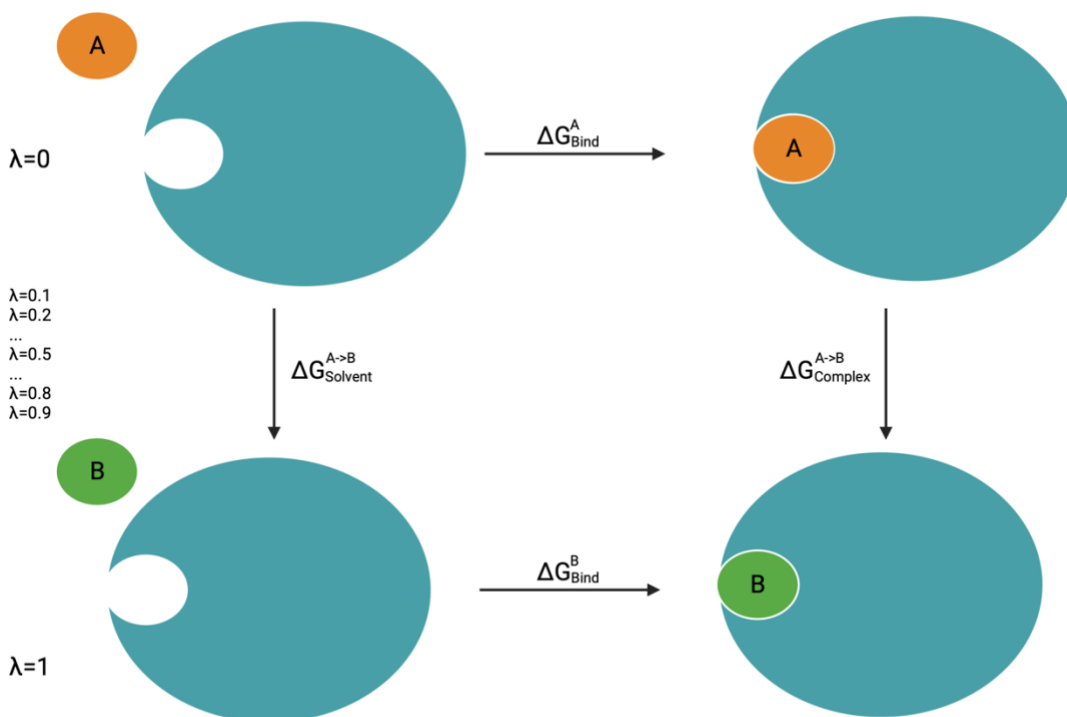
Along each leg of the simulation, the free energy change for converting ligand A into ligand B in both the bound and the unbound form can be estimated using the Zwanzig equation<sup>124</sup>:

$$\Delta G^{A \rightarrow B} = -kT \ln \langle e^{-\frac{(E_B - E_A)}{kT}} \rangle_A$$

where  $k$  is the Boltzmann constant and  $T$  is the temperature of the system. Thus, the free energy change of converting A to B is calculated as the ensemble average of the Boltzmann weighted difference in potential energies between state B and state A, over the trajectory of A. Modern FEP implementations, such as Schrodinger's FEP+, sample from trajectories of both A and B and integrate the results using the Multistate Bennett Acceptance ratio<sup>125</sup>. To facilitate the numerical convergence of this simulation, in practice, the transformation from A to B is performed in a series of alchemical steps denoted by a  $\lambda$  parameter which varies the Hamiltonian of the system using the following formula:

$$\mathcal{H} = \mathcal{H}_0 + \lambda \mathcal{H}_A + (1 - \lambda) \mathcal{H}_B$$

where  $\mathcal{H}_A$  represents the Hamiltonian of atoms that are representative of the initial protein-ligand A complex,  $\mathcal{H}_B$  similarly represents the Hamiltonian of the protein-ligand B complex and  $\mathcal{H}_0$  represents the Hamiltonian of remaining atoms in the system which are not transformed over the simulation. By slowly increasing the lambda parameter from 0 to 1, the Hamiltonian is gradually shifted from describing ligand A to ligand B. The intermediate alchemical states do not correspond to any physically realizable state, however the contributions of these states to the total binding free energy cancel in the thermodynamic cycle, and therefore do not influence the final result beyond smoothing the numerical calculation.



**Figure 3.1 Relative Binding Free Energy Perturbation Thermodynamic Cycle.**

Direct calculation of the change in binding free energy of ligand B (green) relative to ligand A (orange) to a common receptor protein (cyan),  $\Delta\Delta G$ , is recast as the difference in free energy obtained by the alchemical conversion of ligand A to ligand B in complex with the protein (right) compared to converting ligand A to ligand B in bulk solvent (left). This conversion proceeds through a series of alchemical steps controlled by the parameter,  $\lambda$  (left), which gradually converts the Hamiltonian describing the system from ligand A to ligand B.

By converting ligand A to ligand B as described above, Relative Binding FEP (RBFEP) is able to calculate the difference in free energy associated with a protein binding ligand A compared to binding ligand B. To calculate the absolute binding affinity for compound A, compound B can instead be modeled as a dummy ligand whose interactions are completely decoupled from the rest of the system. Under this model, the state where the protein is bound to ligand B and the protein is not bound to ligand B are physically identical, thus  $\Delta G_{\text{Bind}}^{\text{B}} = 0$  and  $\Delta G_{\text{Bind}}^{\text{A}} = \Delta G^{\text{A} \rightarrow \text{B}}_{\text{solvent}} - \Delta G^{\text{A} \rightarrow \text{B}}_{\text{complex}}$ . This approach is called Absolute Binding Free Energy Perturbation (ABFEP). In

practice, positional restraints and additional states in the thermodynamic cycle may be needed to achieve reliable results with ABFEP<sup>126</sup>.

FEP and related alchemical techniques like thermodynamic integration<sup>127</sup> have been used effectively to estimate the ligand solvation free energies<sup>128</sup>, enzyme kinetics<sup>129</sup>, relative binding affinities of small molecule congeneric series<sup>120, 130, 131</sup>, and to predict the effects of mutations on protein-protein binding affinities<sup>132</sup> and in antibody design<sup>133</sup>. In this chapter, we use ABFEP to estimate the binding affinity of select PCIs of interest. In chapter 5 we use RBFEP to estimate the relative binding affinities for several congeneric series of ligands against a shared protein target, the *Machilis hrabei* olfactory receptor 5 (MhOR5) as well as to assess the effect of mutating MhOR5 binding pocket residues on the binding affinity to the small molecule, eugenol.

## **3.2 Results**

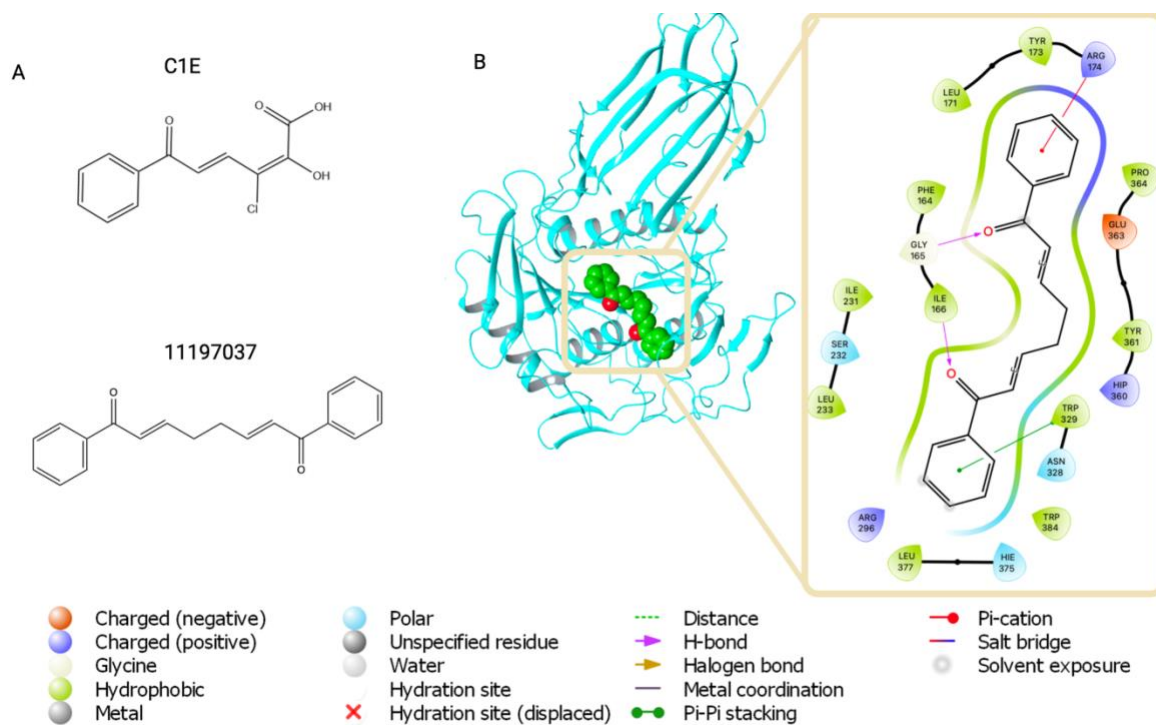
### **3.2.1 Lead compound discovery**

One area in which we expect PrePCI is particularly well-suited is in the identification of novel lead compounds for therapeutic targets of interest. These lead compounds may initially have low affinity and lack complex modifications which may confer selectivity but provide an initial scaffold which can be optimized to inhibit specific protein targets of interest.

#### **3.2.1.1 Lead compound discovery – ACOT4**

We used PrePCI to search for compounds predicted to bind to peroxisomal acyl-coenzyme A thioesterase 4 (ACOT4), a lipid metabolism enzyme which catalyzes the cleavage of acyl-coenzyme A to coenzyme-A (CoA) and free fatty acids. A recent study found that pancreatic ductal adenocarcinoma cells are dependent on free CoA generated by ACOT4 while knockdown and

catalytic inactivation of ACOT4 impaired tumor formation and proliferation, suggesting ACOT4 as a possible therapeutic target<sup>134</sup>. Notably, while no PCIs are predicted for ACOT4 on the basis of sequence similarity, structural alignment identified 41 PDB compounds with LT-scanner scores  $\geq 0.3$ . We chose to focus on C1E (Figure 3.2.A) due to its relatively high LT-scanner score (0.39) and the diversity of similar compounds in PubChem for screening (80 with TC <0.7). The C1E-ACOT4 interaction was predicted based on the crystal structure of C1E complexed with a *Burkholderia xenovorans* C-C hydrolase (PDB ID 2RHT, Chain A). Glide<sup>56, 57</sup> was used to dock C1E into both ACOT4 and the template protein structure as a control, which yielded favorable glide scores of -7.1 and -10.1 kcal/mol, respectively, indicating C1E is a reasonable lead for ACOT4. The 80 similar compounds were similarly analyzed using Glide and among the best scoring ligands was Pubchem CID 11197037 (Figure 3.2) which Glide predicted to bind ACOT4 with an affinity of -9.2kcal/mol. The predicted binding mode of CID 11197037 positions a benzaldehyde ring in a pose similar to the template while the remainder of the compound provides additional contacts and more fully occupies ACOT4's active site. We subsequently performed IFD-MD to optimize the predicted ACOT4-small molecule interaction model and used AB-FEP to estimate the binding affinity of the interaction which yielded estimated binding affinity of -9.4 kcal/mol. We anticipate further refinement of 11197037 will enable the identification of more potent ACOT4 binders.



### Figure 3.2: Lead compound discovery for ACOT4

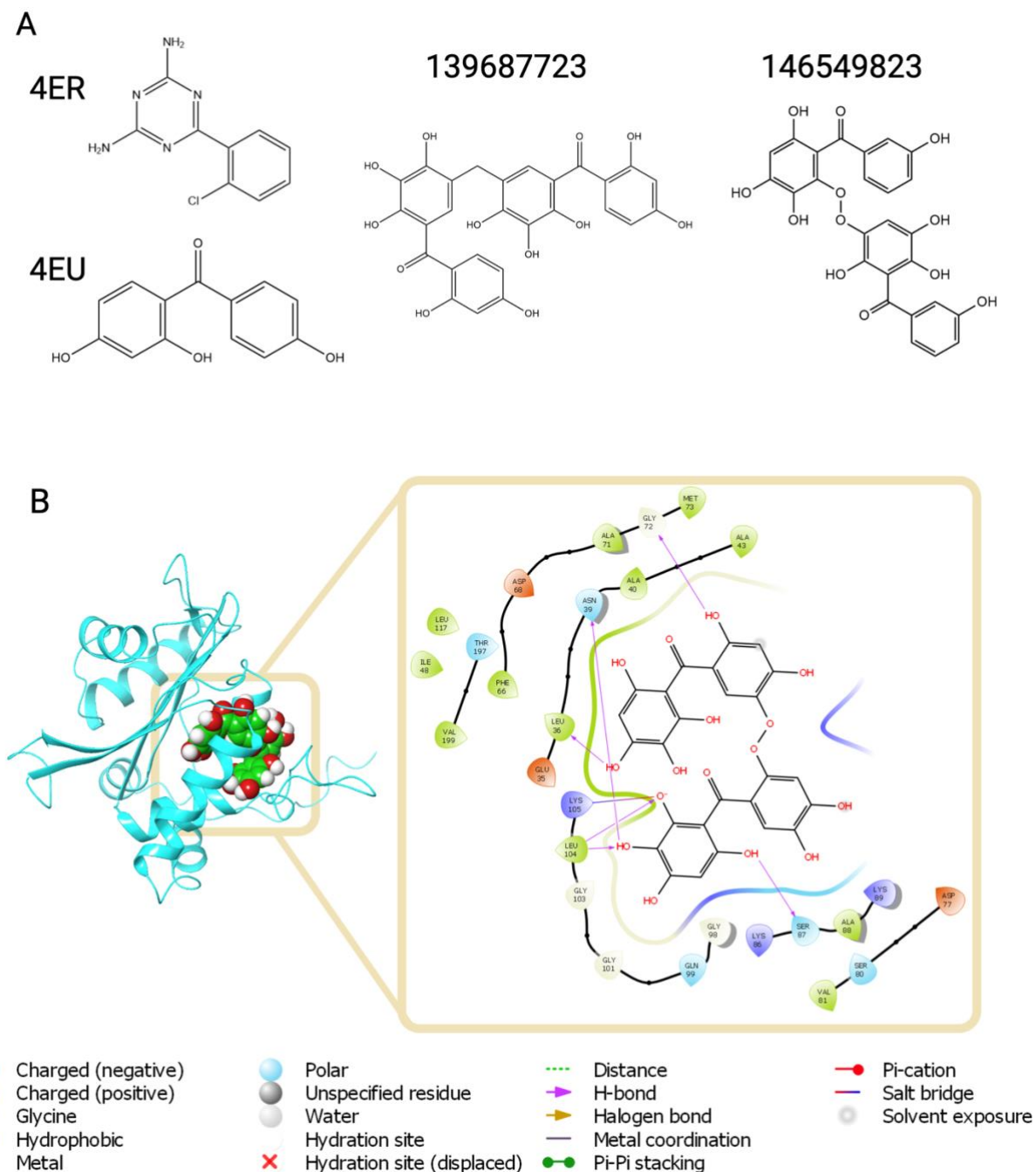
PrePCI guided structure-based virtual screening. (A) PrePCI predicts 41 PDB compounds bind to ACOT4 with LT-scanner score  $\geq 0.3$ , including C1E (top). In silico screening of C1E and similar compounds identifies PubChem CID 11197037 as a possible binder (bottom). (B) The docking pose (left) is depicted as a cyan backbone ribbon for the target (ACOT4) and space-filling representation for the compound (11197037). The diagram (right) highlights atomic interactions predicted by docking.

#### 3.2.1.2 Lead compound discovery – MORC2

We used a similar strategy to identify lead compounds for Microrchidia 2 (MORC2), an ATPase involved in epigenetic silencing and transcriptional regulation<sup>135</sup>, whose aberrant activity has been implicated in hepatic and gastric malignancies<sup>136</sup> as well as neurodegenerative conditions including Charcot-Marie Tooth Disease<sup>135</sup>. The ATPase domain of MORC2 is required for the epigenetic suppression of Hippo signaling, which has been shown to play critical roles in inhibiting hepatocellular carcinoma (HCC)<sup>137</sup>. Therefore, MORC2-ATPase inhibitors, which may serve to reactivate Hippo signaling, have been suggested as a novel approach for HCC therapy<sup>137</sup>. PrePCI

predictions for MORC2 can be leveraged as structural models for more in-depth computational screening and optimization.

PrePCI/DB provides 62 novel compounds with LT-scanner scores  $\geq 0.6$  for MORC2, none of which is predicted by sequence similarity and, thus, constitute predictions based solely on structural evidence. Among the top scoring compounds are the drug-like PDB ligands 4ER (PubChem CID 97399) and 4EU (PubChem CID 73852) (Figure 3.3.A). The LT-scanner models for MORC2/4ER and MORC2/4EU provide a springboard for computational chemical screening of 1,175 PrePCI identified compounds with the docking software Glide<sup>56,57</sup>. PubChem compound 139687723 was predicted as the highest affinity ligand with a GlideXP score of  $-10.0$  kcal/mol, which exceeds the predicted affinity with ATP, MORC2's natural ligand ( $-9$  kcal/mol). A second round of screening using compounds retrieved from PubChem that are structurally similar to 139687723 yielded a compound (PubChem ID 146549823) with a predicted Glide binding affinity of  $-15$  kcal/mol. The docked complex, depicted in Figure 3.3.B, highlights the five hydrogen bonds and the salt bridge between the compound and MORC2, and provides a model for further pharmacologic optimization.



**Figure 3.3: Lead compound discovery for MORC2.**

A) Chemical structures of PDB template compounds 4ER and 4EU (left), Pubchem compound 139687723 (center) and Pubchem compound 146549823 (right). B) Predicted three dimensional (left) and two-dimensional interaction diagram of the MORC2-146549823 interaction.

### 3.2.2 Study of protein-metabolite interactions for protein functional annotation

Many peripheral membrane proteins transiently associate with membrane surfaces by recognizing phosphatidylinositol phosphates, such as PI(4,5)P2 (PIP2) and PI(3,4,5)P3 (PIP3)<sup>29, 30, 138</sup>. Structural studies have elucidated the binding mode of a wide array of protein domains to the head groups of PIP2 and PIP3 (denoted here by their PDB IDs, I3P, and 4IP), which appear as non-covalent ligands in 46 and 27 PDB complexes, respectively. PrePCI predicts over 400 targets for each (LT-scanner score of at least 0.3), many of which are novel. Figure 2.1 illustrates an example of a novel I3P target. Paladin was previously annotated as an inactive protein phosphatase but is predicted by PrePCI to bind to I3P through its first protein-tyrosine phosphatase-like domain, suggesting Paladin may be a lipid, rather than protein, phosphatase. Consistent with this prediction, Paladin was recently identified as a PIP2 phosphatase through a colorimetric screen for phosphate in the presence of PIP2<sup>139</sup>. As depicted in Figure 2.1, PrePCI also predicts that Paladin binds Ins(1,4)P2 (PubChem CID 439444), a compound chemically similar to I3P which corresponds to the head group of PI4P. Since PI4P is the product of 5-phosphatase activity against PI(4,5)P2, the prediction that Paladin binds the head groups of both reactant (PI(4,5)P2) and product (PI4P) suggests that Paladin may be a 5-phosphatase. As a cautionary note, PrePCI predicts that Paladin also binds 4IP (Ins(1,3,4,5)P4) albeit with a lower LR (93 for 4IP vs. 315 for 3IP). Additional computational and experimental analysis is required to determine whether Paladin is a PIP2 phosphatase, a PIP3 phosphatase, or both.

The integration of PrePCI with high-throughput lipidomic assays provides structural annotation of protein–lipid interactions, boosts confidence in the discovery of novel binders and, thus, expands phosphatidylinositol phosphate interactomes. Two studies used mass spectrometry-based methods to identify PIP3-binding proteins in HeLa cells<sup>140</sup> and human platelets<sup>141</sup>. In both

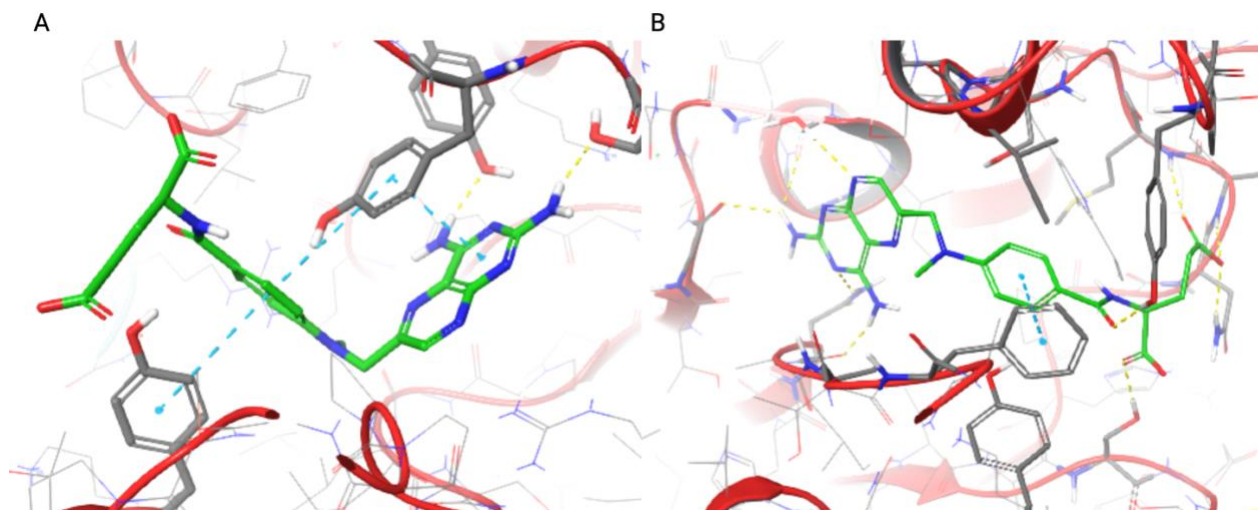
cases, PrePCI predicts 45% of the 30 highest scoring and 25% of all PIP3 binders as 4IP targets (Table 3.S1). Of 21 proteins annotated as known binders<sup>140</sup>, PrePCI predicts all 21 with LRs ranging from 100 to 690 K. Mass spectrometry and PrePCI jointly identify 16 of the known PIP3 binders as well as an additional 70 potentially novel PIP3 interactors (Table 3.S1).

### 3.2.3 Elucidation of drug mechanism of action

The DeMAND algorithm is a network-based approach to elucidating drug mechanism of action as defined by a compound's direct and indirect targets (effectors and modulators) through the analysis of cellular perturbation gene expression profiles<sup>142</sup>. The integration of DeMAND and PrePCI predictions can identify direct targets and off-targets of compounds within particular cellular contexts on a genome-wide scale. For example, high-scoring predictions in both DeMAND and PrePCI for methotrexate (a chemotherapy agent and immune-system suppressant) and genistein (a flavonoid in clinical trials as a treatment for prostate cancer) recapitulate known targets and highlight potential off-targets in diffuse large B cell lymphoma cells (Table 3.S2).

The WW domain-containing oxidoreductase (WWOX) is predicted as a novel target of methotrexate. WWOX has been shown to regulate susceptibility of squamous cell carcinoma to methotrexate, and small interfering RNA against WWOX blocked methotrexate-mediated cell death<sup>143</sup> supporting the plausibility of WWOX as a direct target. Following the same strategy that was used for ACOT4, we performed rigid body docking of methotrexate into our homology model for WWOX to obtain an initial pose and docking energy estimate. This initial pose was subsequently refined using IFD-MD followed by ABFEP yielding a predicted binding free energy of -8.0 kcal/mol, further supporting the possibility that WWOX is a direct target of methotrexate with roughly micromolar affinity (Figure 3.4). Another example is Polo-like kinase 1 (PLK1)

which is predicted as an off-target of genistein, which was shown to function as a mitotic blocker by directly inhibiting PLK1 activity in transformed cells<sup>144</sup> supporting PLK1 as a direct target.



**Figure 3.4 Comparison of template and predicted WWOX-methotrexate binding mode.**

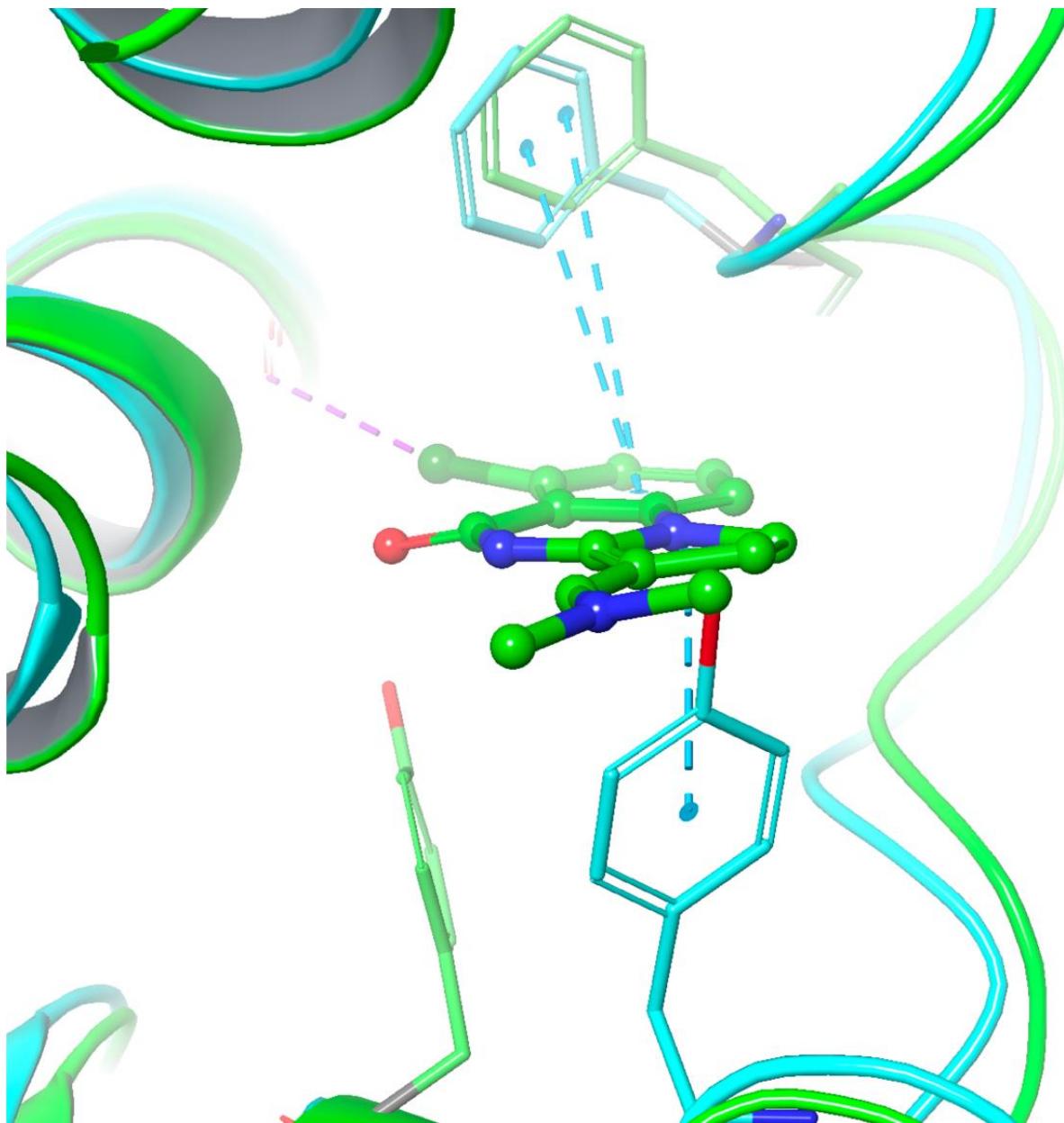
(A) Methotrexate binding pose with *Trypanosoma cruzi* Pteridine Reductase 2 (PDB 1MXF) prepared using Schrodinger Protein Preparation Wizard with default settings. (B) Methotrexate binding pose with human WWOX obtained from ABFEF. One of two aromatic interactions made in the template is preserved in WWOX along with additional hydrogen bonds with terminal hydrophilic groups.

### 3.2.4 Prediction of Ligands binding at Protein-Protein Interaction Interfaces

Additionally, because LT-scanner identifies specific binding sites for each of its predictions, it is possible to evaluate whether two binding sites are predicted to overlap. Of particular interest to our lab is the co-occurrence of ligand and PPI binding sites as such sites could potentially be targeted to disrupt PPIs with small molecules<sup>145</sup>. We performed a preliminary analysis, looking at high confidence LT-scanner predictions (LT-scanner score > 0.4) with limited sequence similarity (e-value > 1) which occur at binding sites composed of at least 5 residues and which, when aligned to a PPI predicted by PrePPI, positions the ligand inside the partner protein (at least 50% of ligand heavy atoms within 2Å of a protein heavy atom). We focused specifically

on cases where the PCI protein was predicted to bind recently defined cancer master regulators<sup>119</sup>, with the hypothesis that disrupting the interaction of these proteins with cancer master regulators might disrupt cancer homeostasis<sup>145</sup>. Table 3.S3 includes all of the 475 PrePCI predictions meeting the above criteria we were able to identify.

One notable example involves the kinase and master regulator BAZ1B. PrePPI predicts that BAZ1B interacts with a protein called PHIP via PHIP's bromodomain while PrePCI predicts that PHIP binds to the compound 5XL via its bromodomain, positioning 5XL within the predicted BAZ1B binding site (Figure 3.5). More rigorous evaluation of this and additional predictions in Table 3.S3 using the techniques described above could enable pharmacological targeting of master regulator proteins by interfering with the proteins they are likely to interact with. While Table 3.S3 outlines only a subset of possible cases, a more systematic analysis comparing LT-scanner binding sites and PrePPI/Predus predicted binding sites could enable lead identification for PPIs of clinical interest.

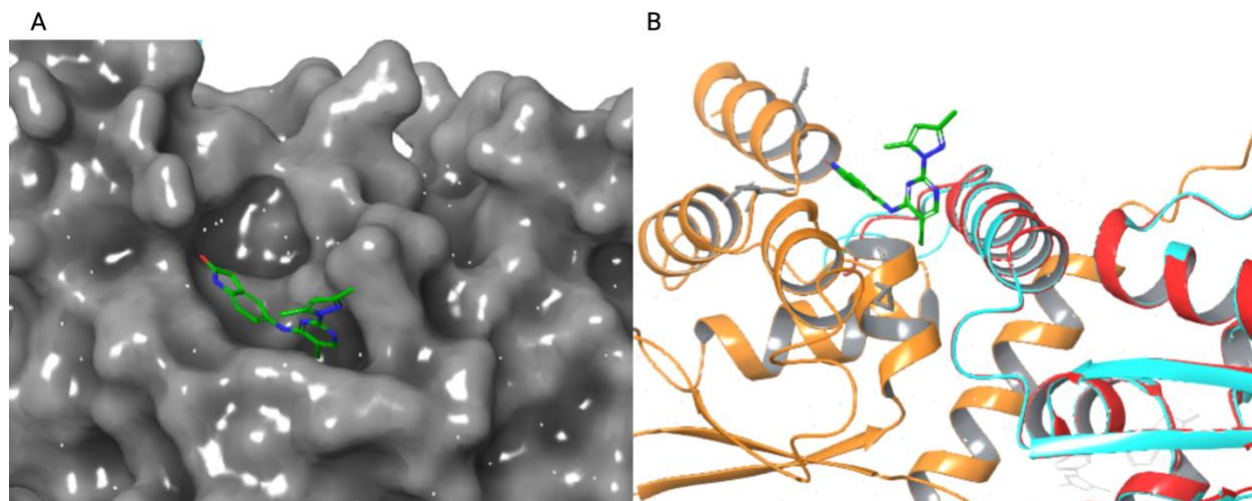


**Figure 3.5 PrePCI interaction model of putative BAZ1B-PHIP interaction disruptor, 5XL in complex with PHIP.**

Unrefined structural model of PHIP (cyan) superimposed on the PB1 bromodomain of PBRM1 in complex with the ligand 5XL (green, PDB ID 5FH7). PHIP conserves interactions made between the template protein, including a pi-pi interaction to structurally conserved phenylalanine residues (upper cyan dashed lines), a halogen hydrogen bond to a backbone carbonyl oxygen (purple dashed line) and is predicted to form an additional pi-pi interaction with a neighboring tyrosine (cyan dashed line).

### 3.3 Limitations of PrePCI structural predictions

Additional limitations of PrePCI's structural predictions of which the user should be aware became apparent during the preparation of the above vignettes. First, because LT-scanner compares the binding interfaces of individual protein chains, it can yield high scoring predictions for compounds which are experimentally observed to bind adjacent to protein-protein interfaces that are present in the template PDB complex. While LT-scanner detects the high structural similarity with one chain in the template, it is currently unable to penalize the query PCI for lacking additional contacts provided by the second protein chain and can thus predict interactions that appear visually implausible in the absence of the partner protein. We observed such a case in the analysis of a prediction for the protein RHOBTB2. In this example, the sequence and model for RHOBTB2 are matched to the PDB complex of the BCL6 BTB-domain bound to PDB compound TJ3 (PubChem CID: 137350052, PDB ID 5mwd, Figure 3.6) with BLAST e-value 0.022 and LT-scanner score 0.84, respectively. Despite the strong LT-scanner score of 0.84, visual inspection of the putative PCI illustrates that roughly half of the small molecule contacts the RHOBTB2 molecule while the remaining half appears exposed to solvent (Figure 3.6B). We speculate that this feature of LT-scanner can be leveraged to identify compounds which can bind at protein-protein interfaces and act as molecular glues, however implementing this within the PrePCI pipeline is left to future work.

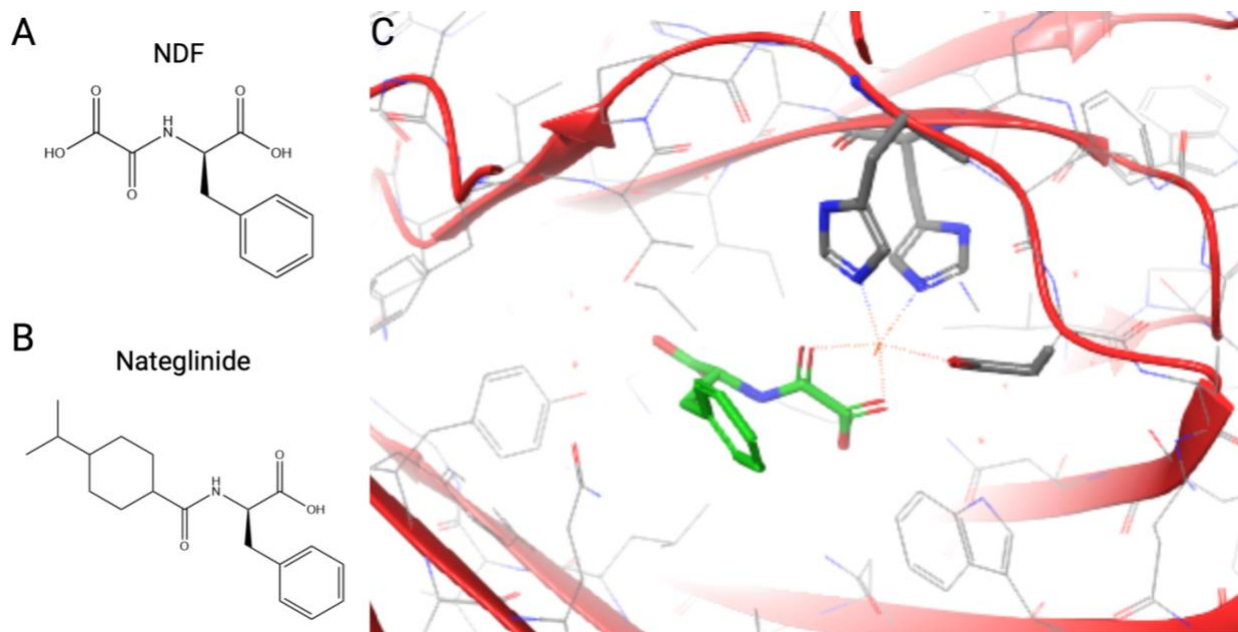


**Figure 3.6 PrePCI predictions at PPI interfaces.**

(A) Surface representation of Bcl6 homodimer (gray surface) in complex with ligand TJ3 (PDB 5MDW). (B) Alignment of RHOBTB2 (cyan) to one of the Bcl6 homodimer chains. This interaction yields a high LT-scanner score (0.84) despite a poor e-value (0.22), suggesting a high degree of structural similarity. However, without additional stabilizing contacts provided by the second chain (orange), half of the small molecule is entirely exposed to solvent instead of buried as in (A). Consequently, it is unlikely that TJ3 binds RHOBTB2 unless RHOBTB2 both forms a heterodimer and shares high structural similarity with Bcl6 at the second binding site (orange chain).

Second, the protein models provided by PrePCI lack metal ions, cofactors and compounds which can play central roles in ligand binding. Accordingly, subsequent analysis without accounting for the presence of these ions can lead to misleading results. We observed such a case when searching for drug repurposing opportunities for the anti-diabetes drug, nateglinide (PubChem CID: 5311309), and retrieved ALKBH4, a dioxygenase involved in the demethylation of actin and the promotion of cell migration<sup>146</sup>. ALKBH4 has recently been shown to promote tumorigenesis in non-small cell lung cancer<sup>147</sup>, and its knockdown induces G1 cell cycle arrest and abrogates proliferation, suggesting repurposing of nateglinide or its optimized derivatives as a possible therapeutic strategy for treating NSCLC. The prediction for ALKBH4/nateglinide arises from the chemical similarity between nateglinide and the PDB compound NDF (TC = 0.52; Figure

3.7A) which was predicted to bind ALKBH4 by LT-scanner (score = 0.29) based on the similarity between ALKBH4 and the template complex HIF1AN/NDF (PDB ID: 1yci, chain A). Moreover, LT-scanner predicts that nateglinide would bind directly to key catalytic residues in ALKBH4's active site<sup>147</sup> and could therefore act as an inhibitor. Docking simulations with Glide in the absence of the bound iron ion predicted a modest affinity of NDF for HIF1AN (score -5.5 kcal/mol) and a stronger affinity of nateglinide for ALKBH4 (score -6.3 kcal/mol) which was increased upon IFD-MD (score -8.5 kcal/mol). Nevertheless, visual inspection of the binding mode of the template ligand, NDF, with the template protein, HIF1AN, showed that NDF bound to HIF1AN by coordinating a bound iron ion using its terminal carbonyl and carboxylate groups. Nateglinide in contrast, lacks a terminal carboxylate group and therefore is less likely to be a true interactor as it would be unable to coordinate ALKBH4's active site iron ion in the same manner as NDF. This case underscores the importance of carefully reviewing predicted interactions for metal ions or cofactors which may be absent from the query model as well as confirming that essential chemical warheads in the reference compound are preserved in the predicted compound.



**Figure 3.7 PrePCI predictions lack metal ions.**

Chemical structures of (A) NDF and (B) Nateglinide, both of which were predicted to bind to ALKBH4 by PrePCI. (C) PDB Structure 1YCI, displaying the coordination of a bound iron ion via NDF's (green) carbonyl and carboxylate groups. Analogous groups are missing from Nateglinide, undermining the prediction that nateglinide binds ALKBH4.

Third, the Tanimoto coefficient measures overall topological similarity and may therefore identify compounds lacking critical interacting functional groups or with large changes in physical features such as net charge. The user is therefore encouraged to review the template PDB complex on which the prediction was based to verify whether the query compound lacks features which could make the predicted interaction more or less plausible.

**3.4 Discussion**

We have presented the PrePCI algorithm, and a corresponding database PrePCI/DB, which integrates protein structure and chemical and sequence similarity to predict protein compound interactions (Figure 2.1). PrePCI is an extension of our template-based PCI prediction algorithm,

LT-scanner<sup>91</sup>, which identifies protein targets of small molecules present in the PDB<sup>94</sup>. The LT-scanner query model database has been updated with structures from the AlphaFold Protein Structure Database<sup>102, 103</sup> providing essentially complete structural coverage for all human protein-coding genes. In addition, PrePCI uses chemical compound similarity based on Tanimoto coefficients among chemical fingerprints to link PDB compounds to compounds in PubChem, which has the effect of increasing the number of compounds that can be explored by more than 200-fold. The increased structural coverage of the human proteome and the expanded chemical space have enabled the evaluation of over 5 billion protein-compound pairs, each with component scores and an overall likelihood ratio that allow users to prioritize predictions. While this constitutes a significant expansion over our previous work, it still excludes many compounds which are dissimilar from those in the PDB but for which non-structural bioactivity data is available. Integration of PrePCI scores with machine learning methods could enable further expansion into chemical space while preserving structure-based and proteome-scale predictions. Finally, PrePCI achieves high precision for binary classification on large-scale bioactivity data (Figures 2.2, 2.S1) strongly suggesting that many novel interactions may be present among the most highly ranked predictions in the negative PCI dataset. While PrePCI does not currently attempt to predict the strength of the affinity of the predicted interaction, it may be possible to modify its scoring function to enable affinity prediction. This could be attempted using rapid pose scoring methods like RF-score or graph neural networks to score the interaction models generated by LT-scanner, or alternatively using machine learning to predict corrections to the affinity measurements reported for the template used by LT-scanner based on the degree to which the template binding interface is reproduced by the query protein (i.e. the LT-scanner score) as well as the degree of similarity between the template and query compounds (i.e. Tanimoto Similarity).

We have demonstrated how PrePCI can be used for common medicinal chemistry tasks, such as the identification of small chemical fragments as lead compounds for structure-based virtual screening (Figures 3.2, 3.3) or elucidation of a drug's mechanism of action (Figure 3.4), as well as the generation of biological hypotheses by detecting novel protein-metabolite interactions (Figure 2.1). Importantly, for compounds present in the PDB<sup>94</sup>, PrePCI generates 3D interaction models and predicts interfacial residues. Given how they are constructed, the models are expected to be crude but can be refined with various docking strategies as illustrated above. While an interaction model is not created for a PCI predicted by chemical similarity, the predicted compound can be cross-docked to the template PDB compound binding site in the underlying LT-scanner model. In this regard it is important to emphasize that PrePCI is primarily intended for hypothesis generation. PrePCI/DB, which encompasses scores for 5 billion protein-compound pairs involving 6.8 million compounds and 19,797 protein targets, is a conveniently accessible structure-informed resource to search for compounds that potentially bind a given protein or, alternatively, proteins that are potential targets of a given compound.

### **3.5 Methods**

#### **Rigid Body Docking**

An initial putative interaction model of each PCI described above was created by structurally aligning the query protein to the template protein using *ska*<sup>97</sup>. These files are easily downloadable directly from the PrePCI website. Structures of both the template protein and the query protein were prepared in the presence of the template ligand using the Protein Preparation Wizard in Maestro version 13.1 using default settings with the following modifications.

- Converge heavy atoms to RMSD 1A (increased from 0.3A)
- Cap termini
- Fill in missing chains (attempts to model any missing residues in the protein structure using Schrodinger's PRIME software)

Receptor grids representing the pharmacophoric properties of the binding surface were generated around the template ligand, setting all neighboring groups as rotatable with other settings taken as the defaults.

Ligand structures were prepared using Ligprep: PDB ligands were prepared from directly from their coordinates and chirality was inferred from 3D structure; screening compounds were prepared from SMILES strings and all combinations of chiral centers were generated to expand diversity of the screening pool. All docking was performed with flexible ligand sampling using the XP scoring function<sup>56, 57</sup>.

#### IFD-MD

An interaction model for the top-ranking Glide pose for each PCI was created by merging each protein structure with the ligand pose. This complex model was then used as an initial structure for non-covalent IFD-MD, including running Forcefield Builder to compute missing forcefield parameters for the ligand but otherwise using default settings.<sup>58</sup>

#### AB-FEP

The top ranked pose generated by IFD-MD for each protein-compound pair was subsequently imported in to the FEP+ graphical user interface in Maestro. "Absolute" was chosen for the "Calculate Binding Free Energy for:" setting. Simulations were run with 150mM NaCl and otherwise default settings, using a  $\mu$ VT ensemble with 5ns simulation time and 1ns MD simulation time.<sup>120, 130</sup>

### 3.5 Supplemental Information

A

<i>Uniprot</i>	<i>Gene</i>	<i>Score</i>	<i>LT-scanner</i>	<i>BLAST evalue</i>	<i>Total LR</i>
<b>Q14644</b>	<b>RASA3</b>	30.51584	0.39	1.00E-11	128.1
O15530	PDPK1	28.95214	1	2.00E-87	1,082,187.3
<b>Q86UU1</b>	<b>PHLDB1</b>	27.537	0.9	0.0004	22,278.0
<b>Q15283</b>	<b>RASA2</b>	26.21455	0.36	3.00E-13	128.1
Q8WWN8	ARAP3	23.20361	0.34	0.007	116.1
<b>Q86SQ0</b>	<b>PHLDB2</b>	23.17382	0.49	0.000002	303.7
Q12965	MYO1E	21.90983	0.38	NA	18.3
O43739	CYTH3	21.6186	0.72	0	4,216,057.6
Q14185	DOCK1	20.11523	0.29	NA	10.3
Q15438	CYTH1	19.31744	0.82	0	8,273,341.1
Q92766	RREB1	18.87752	0.23	NA	10.3
Q8TCU6	PREX1	17.36272	0.44	9.00E-96	25,229.0
<b>Q9UQC2</b>	<b>GAB2</b>	16.15195	0.33	0.00001	116.1
<b>Q15057</b>	<b>ACAP2</b>	15.44187	0.36	3E-07	116.1
Q8N110	DOCK4	14.74634	0.3	NA	18.3
Q53GA4	PHLDA2	14.54022	0.25	NA	10.3
<b>Q13480</b>	<b>GAB1</b>	13.88568	0.35	0.00002	116.1
Q9BZ29	DOCK9	13.69255	0.31	0.003	116.1
Q9NYT0	PLEK2	13.55231	0.51	0.000003	1,281.0
O00159	MYO1C	12.1674	0.35	NA	18.3
P42680	TEC	11.95643	0.49	2.00E-32	1,408.3
Q9HB19	PLEKHA2	11.39867	0.47	5E-08	303.7
<b>Q9HD67</b>	<b>MYO10</b>	10.606	0.43	1E-07	303.7
Q9HB21	PLEKHA1	9.810463	0.53	3.00E-13	1,413.1
Q9UPU7	TBC1D2B	9.754505	0.37	NA	18.3
<b>Q9UH65</b>	<b>SWAP70</b>	9.55796	0.38	0.000007	116.1
Q99418	CYTH2	9.0366	0.59	3.00E-179	475,732.4
P98082	DAB2	8.734033	0.36	NA	18.3
Q86WG5	SBF2	7.7552	0.43	0.0001	303.7
Q8WZ64	ARAP2	7.626003	0.36	0.0001	116.1
O94875	SORBS2	7.359731	0.21	NA	10.3
O43795	MYO1B	6.119768	0.34	NA	18.3
Q96P48	ARAP1	5.860523	0.33	0.00001	116.1
Q9UBS4	DNAJB11	5.825011	0.2	NA	3.3
Q01082	SPTBN1	5.094771	0.28	0.16	65.4
Q8N8S7	ENAH	5.023512	0.34	NA	18.3

<b>P31751</b>	<b>AKT2</b>	4.28438	0.39	5.00E-54	714.0
<i>Q96PK6</i>	RBM14	3.925916	0.2	NA	3.3
<i>Q9HAU0</i>	PLEKHA5	3.571351	0.31	3.00E-32	538.2
<b>O95248</b>	<b>SBF1</b>	3.277	0.37	0.00002	116.1
<i>O60942</i>	RNGTT	3.154942	0.31	NA	18.3
<i>Q9Y6Q9</i>	NCOA3	3.143484	0.22	NA	10.3
<i>Q06124</i>	PTPN11	2.737641	0.33	NA	18.3
<i>Q13112</i>	CHAF1B	2.581469	0.29	NA	10.3

## B

<i>Uniprot</i>	<i>Gene</i>	<i>Score</i>	<i>LT-scanner</i>	<i>BLAST evalue</i>	<i>Total LR</i>
<i>Q06187</i>	BTK	3109.1	0.89	3.00E-98	1,850,436.9
<b>Q14644</b>	<b>RASA3</b>	2083.6	0.39	1.00E-11	128.1
<b>Q9UN19</b>	<b>DAPP1</b>	777.1	1	7.00E-73	687,828.5
<i>Q9Y2L6</i>	FRMD4B	554.8	0.28	NA	10.3
<i>P42680</i>	TEC	522.7	0.49	2.00E-32	1,408.3
<i>Q99418</i>	CYTH2	384.3	0.59	3.00E-179	475,732.4
<i>Q15438</i>	CYTH1	171.5	0.82	0.00E+00	8,273,341.1
<b>Q15027</b>	<b>ACAP1</b>	133.8	0.41	3.00E-05	303.7
<i>O43739</i>	CYTH3	132.1	0.72	0.00E+00	4,216,057.6
<i>Q92556</i>	ELMO1	131.2	0.21	NA	10.3
<b>Q15283</b>	<b>RASA2</b>	131	0.36	3.00E-13	128.1
<i>Q8WXE9</i>	STON2	115.6	0.31	NA	18.3
<i>Q86UU1</i>	PHLDB1	106.9	0.9	4.00E-04	22,278.0
<i>Q14185</i>	DOCK1	75.8	0.29	NA	10.3
<i>Q8WWN9</i>	IPCEF1	69.7	0.25	5.00E-10	72.1
<i>O15530</i>	PDPK1	46.3	1	2.00E-87	1,082,187.3
<i>Q9HB19</i>	PLEKHA2	45.2	0.47	5.00E-08	303.7
<b>O75689</b>	<b>ADAP1</b>	42.1	0.46	4.00E-03	303.7
<b>Q15057</b>	<b>ACAP2</b>	40.3	0.36	3.00E-07	116.1
<b>Q8WWW8</b>	<b>GAB3</b>	37.1	0.37	1.00E-07	116.1
<b>Q9UQC2</b>	<b>GAB2</b>	34	0.33	1.00E-05	116.1
<i>B011T2</i>	MYO1G	24.8	0.27	NA	10.3
<i>O00159</i>	MYO1C	15.6	0.35	NA	18.3
<i>Q9UPU7</i>	TBC1D2B	15.6	0.37	NA	18.3
<i>P49757</i>	NUMB	15	0.32	NA	18.3
<i>Q96P48</i>	ARAP1	14.8	0.33	NA	18.3
<i>Q04837</i>	SSBP1	10.7	0.21	NA	10.3
<b>P31751</b>	<b>AKT2</b>	10.2	0.39	5.00E-54	714.0
<i>P98082</i>	DAB2	10.1	0.36	NA	18.3

<i>Q12965</i>	MYO1E	10	0.38	NA	18.3
<i>O75791</i>	GRAP2	9.9	0.23	NA	10.3
<i>O00160</i>	MYO1F	8.9	0.28	NA	10.3
<i>Q8NEU8</i>	APPL2	7.8	0.26	4.00E-02	65.4
<i>Q9HB21</i>	PLEKHA1	7.4	0.53	3.00E-13	1,413.1
<i>O75563</i>	SKAP2	7.3	0.33	NA	18.3
<i>Q14155</i>	ARHGEF7	6.8	0.21	NA	10.3
<i>Q5JSH3</i>	WDR44	6.5	0.24	NA	10.3
<i>Q13884</i>	SNTB1	5.9	0.26	7.80E-01	65.4
<b><i>Q13480</i></b>	<b>GAB1</b>	5.4	0.35	2.00E-05	116.1
<i>O00560</i>	SDCBP	4.6	0.26	NA	10.3
<i>Q05655</i>	PRKCD	4.6	0.32	NA	18.3
<i>Q01968</i>	OCRL	4.5	0.26	NA	10.3
<i>Q92608</i>	DOCK2	4.3	0.23	NA	10.3
<i>Q99961</i>	SH3GL1	4.3	0.25	NA	10.3
<i>Q15052</i>	ARHGEF6	2.9	0.26	2.70E-02	65.4
<i>O60229</i>	KALRN	2.7	0.29	NA	10.3
<i>Q9Y6R0</i>	NUMBL	2.6	0.35	NA	18.3
<i>Q96HC4</i>	PDLIM5	2.3	0.22	NA	10.3
<i>P29122</i>	PCSK6	2	NA	4.80E+00	10.2
<i>Q13613</i>	MTMR1	1.9	0.28	NA	10.3
<i>Q8N4B1</i>	FAM109A	1.8	0.39	NA	18.3
<b><i>P31749</i></b>	<b>AKT1</b>	1.8	0.91	4.00E-71	469,820.8
<i>O60674</i>	JAK2	1.5	0.25	NA	10.3
<i>P18433</i>	PTPRA	1.2	0.3	NA	18.3
<i>P51452</i>	DUSP3	0.8	0.41	NA	47.9
<i>Q9BPZ7</i>	MAPKAP1	0.6	0.25	NA	10.3
<i>B2RTY4</i>	MYO9A	0.6	0.22	NA	10.3
<i>Q96QR8</i>	PURB	0.5	0.24	NA	10.3

**Table 3.S1. PrePCI predicted binders of Ins(1,3,4,5)P4 (PDB id: 4IP).**

Tables 3.S1.A and B provide the intersection between PrePCI predicted 4IP binders with LT-scanner score  $\geq 0.3$  and experimentally observed PI(3,4,5)P3 binders with scores defined in the respective papers<sup>140, 141</sup>. Proteins in bold are known PIP3 binders<sup>140</sup>.

<i>Uniprot</i>	<i>Gene</i>	<i>compound</i>	<i>demand fdr</i>	<i>LT-scanner</i>	<i>BLAST evalue</i>	<i>Total LR</i>
P11586	MTHFD1	methotrexate	9.55E-05	0.59	4.00E-58	14,685.6
P00374	DHFR	methotrexate	1.10E-03	0.7	2.00E-106	1,044,547.9
O00463	TRAF5	methotrexate	2.97E-03	0.35	NA	18.3
Q9NZC7	WVOX	methotrexate	2.86E-02	0.43	NA	47.9
P14174	MIF	methotrexate	3.21E-02	0.7	3.00E-64	297,155.8
O15067	PFAS	methotrexate	3.55E-02	0.34	NA	18.3
Q9UQ13	SHOC2	methotrexate	5.63E-02	0.3	NA	18.3
P06213	INSR	genistein	3.14E-117	0.43	2.00E-10	335.1
P04626	ERBB2	genistein	7.02E-110	0.33	3.00E-08	116.1
Q05513	PRKCZ	genistein	2.34E-107	0.55	6.00E-20	2,270.2
P53350	PLK1	genistein	1.56E-82	0.54	1.00E-30	5,939.4
P48730	CSNK1D	genistein	5.48E-74	0.53	3.00E-06	1,281.0
Q9UHY1	NRBP1	genistein	1.45E-72	0.36	3.00E-07	116.1
O00141	SGK1	genistein	4.41E-68	0.57	2.00E-27	5,939.4
Q96L34	MARK4	genistein	5.26E-65	0.56	4.00E-38	7,607.9
P06493	CDK1	genistein	4.43E-56	0.51	1.00E-24	2,270.2
Q08345	DDR1	genistein	2.16E-47	0.39	8.00E-10	128.1
O14965	AURKA	genistein	8.00E-39	0.57	1.00E-33	5,939.4
P45983	MAPK8	genistein	1.83E-38	0.49	7.00E-23	538.3
Q96KB5	PBK	genistein	2.68E-38	0.36	1.00E-07	116.1
O60285	NUAK1	genistein	5.29E-38	0.54	3.00E-43	7,607.9
P22607	FGFR3	genistein	8.77E-35	0.64	1.00E-11	4,744.0
O15530	PDPK1	genistein	1.24E-33	0.5	2.00E-25	538.3
Q9HBH9	MKMK2	genistein	1.34E-33	0.52	8.00E-34	5,939.4
Q15746	MYLK	genistein	5.33E-33	0.48	1.00E-71	7,950.3
P27361	MAPK3	genistein	7.14E-31	0.51	4.00E-31	5,939.4
P07947	YES1	genistein	4.48E-29	0.33	2.00E-13	128.1
Q05655	PRKCD	genistein	6.40E-29	0.61	2.00E-23	7,621.6
P54760	EPHB4	genistein	1.45E-28	0.43	5.00E-15	335.1
O00506	STK25	genistein	1.23E-26	0.44	1.00E-18	538.3
Q15208	STK38	genistein	5.13E-26	0.55	5.00E-22	2,270.2
Q13523	PRPF4B	genistein	1.08E-24	0.67	2.00E-11	4,744.0
O96017	CHEK2	genistein	3.24E-24	0.64	2.00E-40	25,541.1
P08069	IGF1R	genistein	3.24E-24	0.42	8.00E-07	303.7
Q9UEE5	STK17A	genistein	1.11E-23	0.57	2.00E-69	33,530.5
P51955	NEK2	genistein	4.65E-22	0.33	7.00E-19	205.7
O14757	CHEK1	genistein	6.55E-21	0.65	4.00E-26	7,621.6
P17612	PRKACA	genistein	2.31E-18	0.57	2.00E-27	5,939.4
P33981	TTK	genistein	3.93E-18	0.37	1.00E-14	128.1

<i>Q8IX11</i>	RHOT2	genistein	2.90E-17	0.33	NA	0.4
<i>Q16584</i>	MAP3K11	genistein	6.36E-16	0.4	5.00E-17	128.1
<i>Q00535</i>	CDK5	genistein	1.02E-15	0.57	6.00E-32	5,939.4
<i>Q7KZ17</i>	MARK2	genistein	6.20E-15	0.57	5.00E-39	7,607.9
<i>O75116</i>	ROCK2	genistein	1.66E-14	0.65	5.00E-18	7,621.6
<i>Q96RU8</i>	TRIB1	genistein	7.48E-14	0.33	3.00E-15	128.1
<i>Q96GD4</i>	AURKB	genistein	1.23E-13	0.56	4.00E-28	5,939.4
<i>P07949</i>	RET	genistein	1.28E-12	0.43	7.00E-12	335.1
<i>Q9NQUS</i>	PAK6	genistein	1.76E-12	0.41	3.00E-30	1,408.3
<i>O95819</i>	MAP4K4	genistein	7.43E-10	0.43	3.00E-14	335.1
<i>Q9BXM7</i>	PINK1	genistein	8.24E-10	0.53	4.00E-05	1,281.0
<i>P24941</i>	CDK2	genistein	1.95E-09	0.57	5.00E-34	5,939.4
<i>P23443</i>	RPS6KB1	genistein	2.69E-09	0.62	5.00E-26	7,621.6
<i>Q99986</i>	VRK1	genistein	3.83E-09	0.47	1.50E-01	303.7
<i>Q92630</i>	DYRK2	genistein	3.96E-09	0.72	1.00E-22	20,119.5
<i>O60566</i>	BUB1B	genistein	5.78E-09	0.35	1.90E-01	116.1
<i>P36896</i>	ACVR1B	genistein	3.27E-08	0.52	6.00E-05	1,281.0
<i>Q9BQI3</i>	EIF2AK1	genistein	1.52E-07	0.51	4.00E-09	1,413.1
<i>P19784</i>	CSNK2A2	genistein	7.11E-07	0.85	5.00E-160	6,788,378.8
<i>Q9BZL6</i>	PRKD2	genistein	8.95E-07	0.6	7.00E-38	25,541.1
<i>P21709</i>	EPHA1	genistein	1.81E-06	0.38	3.00E-13	128.1
<i>Q99640</i>	PKMYT1	genistein	1.86E-06	0.46	8.00E-15	335.1
<i>O00311</i>	CDC7	genistein	3.22E-06	0.57	2.00E-04	1,281.0
<i>P52333</i>	JAK3	genistein	4.84E-06	0.64	9.00E-12	4,744.0
<i>Q9Y2H1</i>	STK38L	genistein	1.01E-05	0.61	8.00E-23	7,621.6
<i>Q99759</i>	MAP3K3	genistein	1.04E-05	0.53	5.00E-24	2,270.2
<i>O75369</i>	FLNB	genistein	1.56E-05	0.31	NA	5.6
<i>Q9H2X6</i>	HIPK2	genistein	1.71E-05	0.67	3.00E-19	7,621.6
<i>Q14004</i>	CDK13	genistein	2.19E-05	0.54	1.00E-12	1,413.1
<i>Q14680</i>	MELK	genistein	2.47E-05	0.53	8.00E-38	7,607.9
<i>O14976</i>	GAK	genistein	4.05E-05	0.53	2.00E-13	1,413.1
<i>P09874</i>	PARP1	genistein	4.89E-05	0.56	7.00E-03	1,281.0
<i>Q9UHD2</i>	TBK1	genistein	7.97E-05	0.56	4.00E-18	2,270.2
<i>P21860</i>	ERBB3	genistein	0.000117976	0.33	5.00E-05	116.1
<i>P80192</i>	MAP3K9	genistein	0.00015534	0.4	1.00E-15	128.1
<i>Q9H2K8</i>	TAOK3	genistein	0.000192385	0.51	7.00E-17	1,413.1
<i>Q16539</i>	MAPK14	genistein	0.000272016	0.32	3.00E-22	205.7
<i>O43683</i>	BUB1	genistein	0.000379351	0.57	1.10E-01	1,281.0
<i>O94804</i>	STK10	genistein	0.000386234	0.47	3.00E-20	538.3
<i>P04049</i>	RAF1	genistein	0.000406214	0.36	8.00E-12	128.1
<i>P10275</i>	AR	genistein	0.000548172	0.48	7.00E-16	335.1
<i>Q15418</i>	RPS6KA1	genistein	0.000887364	0.6	2.00E-43	25,541.1

**Table 3.S2. PrePCI predicted targets of methotrexate and genistein.**

The table provides PrePCI predictions for methotrexate and genistein with LT-scanner score  $\geq 0.3$  and DeMAND FDR  $< 0.01$ <sup>142</sup>.

<i>Master Regulator Uniprot</i>	<i>Master Regulator Gene Name</i>	<i>Partner Protein Uniprot</i>	<i>Partner Protein Gene Name</i>	<i>LT-scanner template</i>	<i>LT-scanner Ligand</i>	<i>LT-scanner Score</i>
P25205	MCM3	P33993	MCM7	6txx_7	AGS_7_1001	1
Q13887	KLF5	Q96SW2	CRBN	6hof_H	Y70_H_502	0.940277
Q14498	RBM39	Q96SW2	CRBN	6hof_H	Y70_H_502	0.940277
Q15233	NONO	Q96SW2	CRBN	6hof_H	Y70_H_502	0.940277
Q7Z3K3	POGZ	Q96SW2	CRBN	6hof_H	Y70_H_502	0.940277
Q8WYB5	KAT6B	Q96SW2	CRBN	6hof_H	Y70_H_502	0.940277
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	6fgl_A	UO1_A_1901	0.882033
O75764	TCEA3	P07478	PRSS2	1ppc_E	MID_E_5	0.861668
Q13887	KLF5	Q96SW2	CRBN	4ci2_B	LVY_B_1429	0.85431
Q14498	RBM39	Q96SW2	CRBN	4ci2_B	LVY_B_1429	0.85431
Q15233	NONO	Q96SW2	CRBN	4ci2_B	LVY_B_1429	0.85431
Q7Z3K3	POGZ	Q96SW2	CRBN	4ci2_B	LVY_B_1429	0.85431
Q8WYB5	KAT6B	Q96SW2	CRBN	4ci2_B	LVY_B_1429	0.85431
Q92766	RREB1	Q96SW2	CRBN	4ci2_B	LVY_B_1429	0.85431
O75764	TCEA3	P07478	PRSS2	1pph_E	OZG_E_1	0.84934
O75764	TCEA3	P07478	PRSS2	1aht_H	APA_H_401	0.84399
P41235	HNF4A	Q13133	NR1H3	3fc6_B	LX2_B_1	0.832198
P41235	HNF4A	P55055	NR1H2	3fc6_B	LX2_B_1	0.814316
O75764	TCEA3	P07478	PRSS2	5wi6_D	OGJ_D_304	0.813079
P17024	ZNF20	Q96SW2	CRBN	4v31_B	DUR_B_151	0.807723
Q13887	KLF5	Q96SW2	CRBN	4v31_B	DUR_B_151	0.807723
Q14498	RBM39	Q96SW2	CRBN	4v31_B	DUR_B_151	0.807723
Q15233	NONO	Q96SW2	CRBN	4v31_B	DUR_B_151	0.807723
Q7Z3K3	POGZ	Q96SW2	CRBN	4v31_B	DUR_B_151	0.807723
Q8WYB5	KAT6B	Q96SW2	CRBN	4v31_B	DUR_B_151	0.807723
Q92766	RREB1	Q96SW2	CRBN	4v31_B	DUR_B_151	0.807723
Q13887	KLF5	Q96SW2	CRBN	5hxb_Z	85C_Z_502	0.79406
Q14498	RBM39	Q96SW2	CRBN	5hxb_Z	85C_Z_502	0.79406
Q15233	NONO	Q96SW2	CRBN	5hxb_Z	85C_Z_502	0.79406
Q8WYB5	KAT6B	Q96SW2	CRBN	5hxb_Z	85C_Z_502	0.79406
Q92766	RREB1	Q96SW2	CRBN	5hxb_Z	85C_Z_502	0.79406
O75764	TCEA3	P07478	PRSS2	1y59_T	TL1_T_350	0.792901
P19883	FST	Q9P0G3	KLK14	3mwi_U	B25_U_1244	0.783029
O75764	TCEA3	P07478	PRSS2	1yyy_1	0KV_1_1	0.778473
P19883	FST	Q9P0G3	KLK14	1y5a_T	TL2_T_790	0.774421
O14753	OVOL1	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
P15336	ATF2	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
P17022	ZNF18	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413

<i>P17024</i>	ZNF20	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
<i>P17035</i>	ZNF28	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
<i>P41182</i>	BCL6	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
<i>Q03112</i>	MECOM	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
<i>Q03938</i>	ZNF90	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
<i>Q13887</i>	KLF5	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
<i>Q14498</i>	RBM39	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
<i>Q15233</i>	NONO	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
<i>Q7Z3K3</i>	POGZ	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
<i>Q8N143</i>	BCL6B	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
<i>Q8NHY6</i>	ZFP28	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
<i>Q8WYB5</i>	KAT6B	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
<i>Q92766</i>	RREB1	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
<i>Q96C00</i>	ZBTB9	Q96SW2	CRBN	5oh8_B	ROL_B_202	0.77413
<i>P19883</i>	FST	Q9P0G3	KLK14	2r2w_U	4PG_U_300	0.772797
<i>O75764</i>	TCEA3	P07478	PRSS2	1ad8_H	MDL_H_250	0.768402
<i>O75764</i>	TCEA3	P07478	PRSS2	1gi7_B	120_B_246	0.766439
<i>O75764</i>	TCEA3	P07478	PRSS2	1qhr_B	157_B_500	0.766192
<i>P19883</i>	FST	Q9P0G3	KLK14	3aav_B	A2C_B_1002	0.761366
<i>O75764</i>	TCEA3	P07478	PRSS2	1way_B	L02_B_1248	0.75468
<i>O75764</i>	TCEA3	P07478	PRSS2	4na9_H	1T7_H_301	0.746836
<i>O75764</i>	TCEA3	P07478	PRSS2	1gia_B	135_B_251	0.742033
<i>O75764</i>	TCEA3	P07478	PRSS2	3shc_H	B01_H_3	0.741675
<i>P19883</i>	FST	Q9P0G3	KLK14	1o5d_H	CR9_H_258	0.740184
<i>O75764</i>	TCEA3	P07478	PRSS2	1tbz_H	00Q_H_343	0.738394
<i>O75764</i>	TCEA3	P07478	PRSS2	4uff_H	6V2_H_1248	0.735455
<i>O75764</i>	TCEA3	P07478	PRSS2	1sc8_U	2IN_U_300	0.733436
<i>O75764</i>	TCEA3	P07478	PRSS2	2feq_H	34P_H_1	0.729227
<i>O75764</i>	TCEA3	P07478	PRSS2	1b5g_H	0ZE_H_372	0.728016
<i>O75764</i>	TCEA3	P07478	PRSS2	1ba8_B	0IT_B_1	0.725843
<i>P41235</i>	HNF4A	Q13133	NR1H3	3kfc_D	61X_D_1	0.725692
<i>O75764</i>	TCEA3	P07478	PRSS2	2zgx_H	29U_H_1601	0.721577
<i>O75764</i>	TCEA3	P07478	PRSS2	1w14_U	SH1_U_1300	0.72057
<i>O75764</i>	TCEA3	P07478	PRSS2	4rko_B	0G6_B_301	0.720203
<i>O75764</i>	TCEA3	P07478	PRSS2	3mwi_U	B25_U_1244	0.719013
<i>O75764</i>	TCEA3	P07478	PRSS2	1no9_H	4ND_H_250	0.71849
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	5rjo_A	1P8_A_1501	0.716271
<i>P41235</i>	HNF4A	P55055	NR1H2	3kfc_D	61X_D_1	0.715312
<i>O75764</i>	TCEA3	P07478	PRSS2	4jzf_H	1NL_H_301	0.715099
<i>O75764</i>	TCEA3	P07478	PRSS2	4yes_B	45S_B_704	0.713914

P19883	FST	Q9P0G3	KLK14	1qhr_B	157_B_500	0.710918
075764	TCEA3	P07478	PRSS2	3rml_H	M31_H_1	0.709897
075764	TCEA3	P07478	PRSS2	2zc9_H	22U_H_1501	0.707233
P19883	FST	Q9P0G3	KLK14	4mpu_B	X2A_B_301	0.706594
075764	TCEA3	P07478	PRSS2	2a2x_H	NA9_H_501	0.705902
075764	TCEA3	P07478	PRSS2	2zgb_H	21U_H_1801	0.705324
075764	TCEA3	P07478	PRSS2	5xg4_U	QUE_U_301	0.70298
075764	TCEA3	P07478	PRSS2	4udw_H	N6L_H_1249	0.701844
075764	TCEA3	P07478	PRSS2	4ufd_H	S49_H_1251	0.701489
075764	TCEA3	P07478	PRSS2	4mpu_B	X2A_B_301	0.701435
075764	TCEA3	P07478	PRSS2	1a5g_H	00L_H_372	0.700301
075764	TCEA3	P07478	PRSS2	1mue_B	CDD_B_248	0.699868
075764	TCEA3	P07478	PRSS2	2bmg_B	I1H_B_1246	0.699292
075764	TCEA3	P07478	PRSS2	4ufe_H	3ZD_H_1249	0.698803
P41235	HNF4A	Q13133	NR1H3	3fal_D	LO2_D_1	0.698585
075764	TCEA3	P07478	PRSS2	5a0a_E	JJS_E_1244	0.695296
075764	TCEA3	P07478	PRSS2	1hut_H	0G7_H_1	0.692958
075764	TCEA3	P07478	PRSS2	1oyt_H	FSN_H_501	0.691842
075764	TCEA3	P07478	PRSS2	1ygc_H	905_H_1	0.689087
075764	TCEA3	P07478	PRSS2	1gj9_B	134_B_251	0.688974
075764	TCEA3	P07478	PRSS2	1c4v_2	IH2_2_370	0.687364
075764	TCEA3	P07478	PRSS2	3sw2_B	F11_B_1	0.68555
075764	TCEA3	P07478	PRSS2	4h42_U	11E_U_301	0.68384
075764	TCEA3	P07478	PRSS2	3ldx_H	NLI_H_1	0.683496
075764	TCEA3	P07478	PRSS2	3u9a_H	S33_H_1	0.683214
075764	TCEA3	P07478	PRSS2	2fes_H	3SP_H_1	0.683128
075764	TCEA3	P07478	PRSS2	3rmo_H	S04_H_1	0.682884
P41235	HNF4A	P55055	NR1H2	3fal_D	LO2_D_1	0.682638
P17022	ZNF18	Q7Z2T5	TRMT1L	2ytz_B	SAH_B_1002	0.681764
P19883	FST	Q9P0G3	KLK14	1yyy_1	0KV_1_1	0.681083
075764	TCEA3	P07478	PRSS2	2ec9_H	24X_H_999	0.679971
075764	TCEA3	P07478	PRSS2	1ypl_H	RA8_H_5555	0.679695
075764	TCEA3	P07478	PRSS2	1a46_H	00K_H_372	0.679274
075764	TCEA3	P07478	PRSS2	1ktt_B	C02_B_1	0.676854
075764	TCEA3	P07478	PRSS2	2bvr_H	4CP_H_1246	0.675863
075764	TCEA3	P07478	PRSS2	3dux_H	64U_H_901	0.675834
075764	TCEA3	P07478	PRSS2	2fs8_C	C3A_C_997	0.670864
075764	TCEA3	P07478	PRSS2	3k9x_B	MBM_B_1	0.668909
075764	TCEA3	P07478	PRSS2	1hdt_H	0E7_H_1	0.668864
075764	TCEA3	P07478	PRSS2	1w10_U	SJ1_U_1245	0.668735

075764	TCEA3	P07478	PRSS2	3tu7_H	OBM_H_1	0.668241
075764	TCEA3	P07478	PRSS2	2cn0_H	F25_H_1246	0.6682
075764	TCEA3	P07478	PRSS2	2ziq_H	26U_H_701	0.667669
P19883	FST	Q9P0G3	KLK14	4e7r_H	0NW_H_302	0.667441
075764	TCEA3	P07478	PRSS2	1w0z_U	SII_U_1245	0.666969
075764	TCEA3	P07478	PRSS2	1mu6_B	CDA_B_248	0.664522
075764	TCEA3	P07478	PRSS2	3rmn_H	M41_H_1	0.663501
075764	TCEA3	P07478	PRSS2	2zdv_H	37U_H_501	0.660587
075764	TCEA3	P07478	PRSS2	1w11_U	SK1_U_1247	0.658819
075764	TCEA3	P07478	PRSS2	3e0p_B	B3C_B_1	0.658502
075764	TCEA3	P07478	PRSS2	3ens_D	ENS_D_301	0.657621
P19883	FST	Q9P0G3	KLK14	1pph_E	0ZG_E_1	0.657591
075764	TCEA3	P07478	PRSS2	1mu8_B	CDB_B_248	0.657481
075764	TCEA3	P07478	PRSS2	5af9_H	SJR_H_1250	0.657376
075764	TCEA3	P07478	PRSS2	4loy_H	6XS_H_304	0.656931
075764	TCEA3	P07478	PRSS2	3c27_B	DKK_B_5000	0.656298
075764	TCEA3	P07478	PRSS2	1ae8_H	AZL_H_600	0.655922
075764	TCEA3	P07478	PRSS2	1awf_H	GR4_H_1	0.655711
P19883	FST	Q9P0G3	KLK14	2zfp_H	19U_H_801	0.655075
P19883	FST	Q9P0G3	KLK14	3sw2_B	FII_B_1	0.654804
P25205	MCM3	P49736	MCM2	6txx_4	AGS_4_902	0.654317
P33993	MCM7	P49736	MCM2	6txx_4	AGS_4_902	0.654317
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	4yab_A	4CN_A_1103	0.654303
075764	TCEA3	P07478	PRSS2	1vzq_H	SHY_H_1256	0.652157
075764	TCEA3	P07478	PRSS2	1ele_E	0QN_E_256	0.651779
P41235	HNF4A	P55055	NR1H2	1pq9_B	44B_B_2501	0.651328
075764	TCEA3	P07478	PRSS2	1d3q_B	BT2_B_400	0.65082
P19883	FST	Q9P0G3	KLK14	1ygc_H	905_H_1	0.650714
P19883	FST	Q9P0G3	KLK14	3hpt_D	YET_D_2	0.650127
075764	TCEA3	P07478	PRSS2	4jyv_H	1OJ_H_301	0.649561
075764	TCEA3	P07478	PRSS2	3f68_H	91U_H_901	0.649512
075764	TCEA3	P07478	PRSS2	2ank_H	N12_H_501	0.648975
075764	TCEA3	P07478	PRSS2	5jfd_H	2TS_H_301	0.648622
075764	TCEA3	P07478	PRSS2	3e6p_H	DFK_H_301	0.648423
075764	TCEA3	P07478	PRSS2	1tom_H	MIN_H_1	0.647096
075764	TCEA3	P07478	PRSS2	3qto_H	10P_H_1001	0.646841
075764	TCEA3	P07478	PRSS2	2cf9_H	348_H_1247	0.646164
P19883	FST	Q9P0G3	KLK14	2znk_H	31U_H_3001	0.645922
P19883	FST	Q9P0G3	KLK14	3k9x_B	MBM_B_1	0.644444
P41235	HNF4A	P55055	NR1H2	4dk7_C	OKS_C_501	0.644394

075764	TCEA3	P07478	PRSS2	1nzq_H	162_H_248	0.640616
075764	TCEA3	P07478	PRSS2	5a0c_B	JJV_B_1001	0.640481
075764	TCEA3	P07478	PRSS2	2vwm_B	LZI_B_1244	0.637784
075764	TCEA3	P07478	PRSS2	2v3h_H	I25_H_1246	0.637486
075764	TCEA3	P07478	PRSS2	1awh_D	GR3_D_1	0.636708
075764	TCEA3	P07478	PRSS2	1bmm_H	BM2_H_1	0.636597
075764	TCEA3	P07478	PRSS2	1lqd_B	CMI_B_301	0.636593
075764	TCEA3	P07478	PRSS2	2bqw_B	IIE_B_1246	0.636406
075764	TCEA3	P07478	PRSS2	3qtv_H	06P_H_1001	0.636215
075764	TCEA3	P07478	PRSS2	3utu_H	1TS_H_901	0.636182
075764	TCEA3	P07478	PRSS2	2aei_H	03R_H_500	0.636147
075764	TCEA3	P07478	PRSS2	4ax9_H	N5N_H_1246	0.63507
075764	TCEA3	P07478	PRSS2	3uwj_H	TIF_H_302	0.63464
075764	TCEA3	P07478	PRSS2	1uvt_H	I48_H_1	0.634007
075764	TCEA3	P07478	PRSS2	1c4u_2	IH1_2_370	0.633917
076071	CIAO1	P61964	WDR5	5eam_B	5MN_B_401	0.6319
075764	TCEA3	P07478	PRSS2	2zff_H	53U_H_2001	0.631169
075764	TCEA3	P07478	PRSS2	1bmn_H	BM9_H_1	0.629743
P25205	MCM3	Q14566	MCM6	6xtx_4	AGS_4_902	0.629328
075764	TCEA3	P07478	PRSS2	2anm_H	CDO_H_1001	0.628987
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	4hvk_A	1AJ_A_201	0.628817
Q9UIG0	BAZ1B	Q9NRL2	BAZ1A	4ir3_A	1FK_A_2006	0.628088
P19883	FST	Q9P0G3	KLK14	1uvs_H	I11_H_11	0.626391
P19883	FST	Q9P0G3	KLK14	2bz6_H	346_H_1258	0.626322
P19883	FST	Q9P0G3	KLK14	4jyv_H	1OJ_H_301	0.6256
075764	TCEA3	P07478	PRSS2	1lpz_B	CMB_B_301	0.625573
075764	TCEA3	P07478	PRSS2	1ay6_H	1ZV_H_5	0.625213
076071	CIAO1	P61964	WDR5	4ql1_B	35Q_B_401	0.625109
075764	TCEA3	P07478	PRSS2	3rly_H	S29_H_1	0.625052
075764	TCEA3	P07478	PRSS2	2p3u_B	663_B_500	0.623212
075764	TCEA3	P07478	PRSS2	1dan_H	0Z6_H_1	0.621971
P19883	FST	Q9P0G3	KLK14	4btu_F	6XS_F_1246	0.621851
075764	TCEA3	P07478	PRSS2	4yt7_H	4K1_H_301	0.621387
Q9UIG0	BAZ1B	Q9NRL2	BAZ1A	6fgl_A	UO1_A_1901	0.620011
P19883	FST	Q9P0G3	KLK14	4jyu_H	1OK_H_301	0.619474
075764	TCEA3	P07478	PRSS2	3kid_U	2BS_U_1	0.619104
075764	TCEA3	P07478	PRSS2	1qbv_H	PPX_H_907	0.61679
075764	TCEA3	P07478	PRSS2	5a2m_H	WX5_H_1249	0.616609
075764	TCEA3	P07478	PRSS2	1qj1_B	166_B_1248	0.615819
P19883	FST	Q9P0G3	KLK14	3qtv_H	06P_H_1001	0.611313

<i>O43679</i>	LDB2	Q9BYB0	SHANK3	3o5n_E	BR0_E_1	0.609837
<i>O75764</i>	TCEA3	P07478	PRSS2	1w0y_H	771_H_1258	0.60973
<i>P19883</i>	FST	Q9P0G3	KLK14	3e16_B	B4C_B_1	0.609234
<i>P19883</i>	FST	Q9P0G3	KLK14	2v3o_H	I26_H_1246	0.607658
<i>O75764</i>	TCEA3	P07478	PRSS2	4yt6_H	4JY_H_301	0.605947
<i>P19883</i>	FST	Q9P0G3	KLK14	1w10_U	SJ1_U_1245	0.60585
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	3svh_A	KRG_A_2	0.605842
<i>O76071</i>	CIAO1	Q86VZ2	WDR5B	5eam_B	5MN_B_401	0.605499
<i>Q09028</i>	RBBP4	Q86VZ2	WDR5B	5eam_B	5MN_B_401	0.605499
<i>Q15542</i>	TAF5	Q86VZ2	WDR5B	5eam_B	5MN_B_401	0.605499
<i>P25205</i>	MCM3	P33992	MCM5	5v8f_6	AGS_6_1101	0.604988
<i>O75764</i>	TCEA3	P07478	PRSS2	5i46_H	67O_H_301	0.603863
<i>O75764</i>	TCEA3	P07478	PRSS2	4baq_B	M4Z_B_1287	0.601855
<i>P19883</i>	FST	Q9P0G3	KLK14	1tbz_H	00Q_H_343	0.600334
<i>O75764</i>	TCEA3	P07478	PRSS2	1uvs_H	I11_H_11	0.599526
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	5dbm_C	58N_C_1201	0.599456
<i>P19883</i>	FST	Q9P0G3	KLK14	4baq_B	M4Z_B_1287	0.598906
<i>O75764</i>	TCEA3	P07478	PRSS2	1uvu_H	DCH_H_1	0.598472
<i>P19883</i>	FST	Q9P0G3	KLK14	2zi2_H	24U_H_801	0.598419
<i>O75764</i>	TCEA3	P07478	PRSS2	1vj9_U	5IN_U_300	0.597463
<i>P41235</i>	HNF4A	P55055	NR1H2	2acl_H	L05_H_104	0.59713
<i>P19883</i>	FST	Q9P0G3	KLK14	4ng9_H	2KE_H_301	0.59692
<i>P19883</i>	FST	Q9P0G3	KLK14	4uff_H	6V2_H_1248	0.595904
<i>P41235</i>	HNF4A	Q13133	NR1H3	4rak_B	652_B_502	0.59563
<i>P25205</i>	MCM3	Q9UJA3	MCM8	6txx_7	AGS_7_1001	0.595342
<i>Q14566</i>	MCM6	Q9UJA3	MCM8	6txx_7	AGS_7_1001	0.595342
<i>O76071</i>	CIAO1	P61964	WDR5	3smr_B	NP7_B_1000	0.593941
<i>O75764</i>	TCEA3	P07478	PRSS2	1uma_H	IN2_H_600	0.59381
<i>O75764</i>	TCEA3	P07478	PRSS2	4q80_B	2YS_B_305	0.593701
<i>O75764</i>	TCEA3	P07478	PRSS2	1lpk_B	CBB_B_301	0.591442
<i>O75764</i>	TCEA3	P07478	PRSS2	4bak_B	M67_B_1287	0.591175
<i>P19883</i>	FST	Q9P0G3	KLK14	4jzd_H	1NJ_H_301	0.589802
<i>O75764</i>	TCEA3	P07478	PRSS2	3ig6_B	438_B_400	0.589473
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	4hxo_A	1A6_A_201	0.589133
<i>O75764</i>	TCEA3	P07478	PRSS2	3uuz_B	0CB_B_481	0.588097
<i>O75764</i>	TCEA3	P07478	PRSS2	2gde_H	SN3_H_401	0.587301
<i>Q14781</i>	CBX2	Q99549	MPHOSPH8	4x3t_F	45E_F_101	0.586922
<i>O75764</i>	TCEA3	P07478	PRSS2	4btt_B	VYR_B_1246	0.586794
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	5ii1_A	6BL_A_801	0.586654
<i>O75764</i>	TCEA3	P07478	PRSS2	2bq6_B	IIB_B_1246	0.586155

<i>O75764</i>	TCEA3	P07478	PRSS2	1etr_H	MIT_H_1	0.58608
<i>P19883</i>	FST	Q9P0G3	KLK14	4ax9_H	N5N_H_1246	0.585311
<i>P19883</i>	FST	Q9P0G3	KLK14	1qbv_H	PPX_H_907	0.584488
<i>P19883</i>	FST	Q9P0G3	KLK14	3rly_H	S29_H_1	0.584414
<i>O75764</i>	TCEA3	P07478	PRSS2	7est_E	OZ2_E_1	0.584033
<i>P41235</i>	HNF4A	P55055	NR1H2	3ips_B	O90_B_1	0.583921
<i>P19883</i>	FST	Q9P0G3	KLK14	3rmn_H	M41_H_1	0.583423
<i>P25205</i>	MCM3	Q9NXL9	MCM9	6txx_7	AGS_7_1001	0.582208
<i>O75764</i>	TCEA3	P07478	PRSS2	1wss_H	3CB_H_2001	0.582021
<i>O75764</i>	TCEA3	P07478	PRSS2	2est_E	2Z5_E_1	0.581833
<i>P19883</i>	FST	Q9P0G3	KLK14	1mu6_B	CDA_B_248	0.580791
<i>O75764</i>	TCEA3	P07478	PRSS2	4az2_B	9MU_B_1258	0.580329
<i>P19883</i>	FST	Q9P0G3	KLK14	3rm2_H	S00_H_1	0.58023
<i>O75764</i>	TCEA3	P07478	PRSS2	1d3p_B	BT3_B_400	0.577634
<i>P19883</i>	FST	Q9P0G3	KLK14	3ig6_B	438_B_400	0.576519
<i>P63165</i>	SUMO1	Q9UBT2	UBA2	3h9j_D	APC_D_359	0.576166
<i>P19883</i>	FST	Q9P0G3	KLK14	1d3p_B	BT3_B_400	0.575389
<i>O75764</i>	TCEA3	P07478	PRSS2	1mmj_N	FR1_N_241	0.572785
<i>O75764</i>	TCEA3	P07478	PRSS2	4bao_B	MVF_B_1287	0.571551
<i>P19883</i>	FST	Q9P0G3	KLK14	3p17_H	99P_H_1001	0.571235
<i>O75764</i>	TCEA3	P07478	PRSS2	2bvxx_H	5CB_H_1246	0.571181
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	4xub_A	43D_A_2001	0.569324
<i>P19883</i>	FST	Q9P0G3	KLK14	2bxu_H	C1D_H_1246	0.566668
<i>P41235</i>	HNF4A	P55055	NR1H2	4dk8_C	0KT_C_501	0.565321
<i>P19883</i>	FST	Q9P0G3	KLK14	1bcu_H	PRL_H_280	0.562643
<i>O75764</i>	TCEA3	P07478	PRSS2	1fpc_H	OZI_H_371	0.56227
<i>P19883</i>	FST	Q9P0G3	KLK14	4isi_H	1GG_H_301	0.561015
<i>P19883</i>	FST	Q9P0G3	KLK14	4nga_H	2KF_H_301	0.559784
<i>P19883</i>	FST	Q9P0G3	KLK14	1fpc_H	OZI_H_371	0.559166
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	4ir4_A	IR4_A_2001	0.557965
<i>P41235</i>	HNF4A	Q13133	NR1H3	1pq9_B	44B_B_2501	0.556996
<i>P19883</i>	FST	Q9P0G3	KLK14	3qwc_H	98P_H_2001	0.556142
<i>P19883</i>	FST	Q9P0G3	KLK14	2a2x_H	NA9_H_501	0.554896
<i>P41235</i>	HNF4A	Q13133	NR1H3	2acl_H	L05_H_104	0.554312
<i>O76071</i>	CIAO1	Q86VZ2	WDR5B	3smr_B	NP7_B_1000	0.554154
<i>Q09028</i>	RBBP4	Q86VZ2	WDR5B	3smr_B	NP7_B_1000	0.554154
<i>Q15542</i>	TAF5	Q86VZ2	WDR5B	3smr_B	NP7_B_1000	0.554154
<i>P19883</i>	FST	Q9P0G3	KLK14	1w12_U	SL1_U_1245	0.553825
<i>P23246</i>	SFPQ	P62736	ACTA2	1yxq_B	SWI_B_600	0.553033
<i>O75764</i>	TCEA3	P07478	PRSS2	2bvs_H	2CE_H_1246	0.55253

<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	4o7e_A	2RN_A_201	0.550644
<i>P19883</i>	FST	Q9P0G3	KLK14	4gch_G	DMC_G_246	0.550151
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	5cq7_A	53G_A_2001	0.548885
<i>P19883</i>	FST	Q9P0G3	KLK14	4zxy_H	4T1_H_301	0.548218
<i>O75764</i>	TCEA3	P07478	PRSS2	2c8y_B	C3M_B_1252	0.546938
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	5enh_A	5QB_A_1501	0.546709
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	4tte_A	36Z_A_1204	0.545625
<i>P19883</i>	FST	Q9P0G3	KLK14	2pks_C	G44_C_101	0.54507
<i>O75764</i>	TCEA3	P07478	PRSS2	1afe_H	ALZ_H_600	0.544905
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	4xya_A	43S_A_201	0.544753
<i>O75764</i>	TCEA3	P07478	PRSS2	1ghy_H	121_H_246	0.544603
<i>O75764</i>	TCEA3	P07478	PRSS2	4nsy_B	2OY_B_301	0.5424
<i>O75764</i>	TCEA3	P07478	PRSS2	4gch_G	DMC_G_246	0.542338
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	6v1k_A	5SW_A_201	0.540813
<i>P25205</i>	MCM3	Q9UJA3	MCM8	3kb2_B	G3D_B_180	0.539954
<i>Q14566</i>	MCM6	Q9UJA3	MCM8	3kb2_B	G3D_B_180	0.539954
<i>P19883</i>	FST	Q9P0G3	KLK14	1ay6_H	1ZV_H_5	0.539342
<i>O76071</i>	CIAO1	P61964	WDR5	6nm7_B	22L_B_201	0.539197
<i>P63165</i>	SUMO1	Q9UBT2	UBA2	6o83_C	VMX_C_1109	0.539044
<i>P19883</i>	FST	Q9P0G3	KLK14	6eo8_H	2FN_H_307	0.533797
<i>P19883</i>	FST	Q9P0G3	KLK14	3da9_B	44U_B_1	0.531789
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	4lzs_A	L46_A_201	0.531597
<i>P19883</i>	FST	Q9P0G3	KLK14	3utu_H	1TS_H_901	0.531108
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	4hvk_A	1AJ_A_201	0.530247
<i>P19883</i>	FST	Q9P0G3	KLK14	4ayy_B	9MX_B_1258	0.52519
<i>P19883</i>	FST	Q9P0G3	KLK14	2bm2_D	PM2_D_3211	0.52511
<i>P41235</i>	HNF4A	P55055	NR1H2	1p8d_B	CO1_B_109	0.52366
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	5ii1_A	6BL_A_801	0.523527
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	4hvk_A	1AJ_A_201	0.523207
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	4xua_A	43C_A_2005	0.522571
<i>P19883</i>	FST	Q9P0G3	KLK14	2c8y_B	C3M_B_1252	0.522231
<i>P41235</i>	HNF4A	Q13133	NR1H3	4dk8_C	OKT_C_501	0.522014
<i>Q14781</i>	CBX2	Q8N8U2	CDYL2	4x3t_F	45E_F_101	0.521841
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	4o77_A	2RE_A_201	0.520711
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	4meo_A	25V_A_202	0.5195
<i>P19883</i>	FST	Q9P0G3	KLK14	3rm0_H	S54_H_2	0.517858
<i>P19883</i>	FST	Q9P0G3	KLK14	2zp0_H	PI0_H_1001	0.517567
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	4c66_A	H4C_A_1168	0.516901
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	5cp5_A	EB0_A_201	0.516145
<i>O76071</i>	CIAO1	Q86VZ2	WDR5B	4ql1_B	35Q_B_401	0.515799

Q09028	RBBP4	Q86VZ2	WDR5B	4ql1_B	35Q_B_401	0.515799
Q15542	TAF5	Q86VZ2	WDR5B	4ql1_B	35Q_B_401	0.515799
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	4meo_A	25V_A_202	0.512039
P19883	FST	Q9P0G3	KLK14	4zxx_H	4T0_H_301	0.50875
Q9UIG0	BAZ1B	Q8WWQ0	PHIP	4nrb_A	2LX_A_2001	0.507301
O75764	TCEA3	P07478	PRSS2	1w7g_H	MIU_H_300	0.506707
P41235	HNF4A	Q13133	NR1H3	1p8d_B	CO1_B_109	0.504845
P19883	FST	Q9P0G3	KLK14	1sb1_H	165_H_1001	0.502313
Q15542	TAF5	Q9BRX9	WDR83	3smr_B	NP7_B_1000	0.501587
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	4o77_A	2RE_A_201	0.499023
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	4uiz_A	N1D_A_1171	0.497095
O75764	TCEA3	P07478	PRSS2	2c93_B	C4M_B_1251	0.496459
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	4f3i_A	0S6_A_203	0.495609
Q15542	TAF5	Q8TED0	UTP15	4x9d_E	U5P_E_104	0.495567
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	4men_A	25K_A_202	0.495554
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	4uyd_A	V1T_A_1171	0.493596
Q9UIG0	BAZ1B	Q8WWQ0	PHIP	4xub_A	43D_A_2001	0.492926
P19883	FST	Q9P0G3	KLK14	1nzq_H	162_H_248	0.491963
O75764	TCEA3	P07478	PRSS2	2b7d_H	C1B_H_258	0.491646
P19883	FST	Q9P0G3	KLK14	4x8v_H	3Z9_H_301	0.491297
P19883	FST	Q9P0G3	KLK14	2fs9_B	C4A_B_998	0.490631
O75764	TCEA3	P07478	PRSS2	1z6j_H	PY3_H_403	0.490067
Q9UIG0	BAZ1B	Q9NRL2	BAZ1A	3q2f_A	OAM_A_1	0.488552
O76071	CIAO1	Q8NBT0	POC1A	5eam_B	5MN_B_401	0.487637
Q15542	TAF5	Q8NBT0	POC1A	5eam_B	5MN_B_401	0.487637
P19883	FST	Q9P0G3	KLK14	8est_E	GIS_E_269	0.487349
Q9UIG0	BAZ1B	Q9NRL2	BAZ1A	4xua_A	43C_A_2005	0.484701
Q9UIG0	BAZ1B	Q8WWQ0	PHIP	4tte_A	36Z_A_1204	0.481636
Q15542	TAF5	Q9Y297	BTRC	6nm7_B	22L_B_201	0.481412
Q14781	CBX2	Q9HC52	CBX8	4x3t_F	45E_F_101	0.480736
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	5d3n_A	L40_A_201	0.479948
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	5fh6_A	5XM_A_801	0.479704
P19883	FST	Q9P0G3	KLK14	1z6j_H	PY3_H_403	0.479436
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	4tz8_A	39U_A_1204	0.478857
Q9UIG0	BAZ1B	Q8WWQ0	PHIP	5dkc_A	5BW_A_1502	0.477989
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	6in1_A	LOC_A_200	0.476522
Q15542	TAF5	Q9BRX9	WDR83	5eam_B	5MN_B_401	0.475927
Q9UIG0	BAZ1B	Q9NRL2	BAZ1A	5u9n_A	BMF_A_501	0.47592
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	5hm0_A	62V_A_201	0.473302
Q15542	TAF5	Q9BZK7	TBL1XR1	6nm7_B	22L_B_201	0.472166

<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	5fh8_A	5XK_A_801	0.472118
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	4rvr_A	3WQ_A_2001	0.472
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	5fh6_A	5XM_A_801	0.470573
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	4e96_A	0NS_A_201	0.469584
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	4ir5_A	IR5_A_2013	0.469107
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	5ii1_A	6BL_A_801	0.467095
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	4ir6_A	IR6_A_2007	0.466729
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	5fh8_D	5XK_D_801	0.466127
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	5dkd_A	5BW_A_1605	0.464858
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	5fh7_A	5XL_A_803	0.463627
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	5dfd_A	59E_A_501	0.46319
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	4meq_A	25O_A_202	0.461558
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	4q0n_E	2XD_E_803	0.461173
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	4ir5_A	IR5_A_2013	0.460274
<i>O75764</i>	TCEA3	P07478	PRSS2	1wun_H	P5B_H_2001	0.460157
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	4q0o_A	2XC_A_802	0.45928
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	5i89_A	69B_A_1203	0.457307
<i>O76071</i>	CIAO1	Q8NBT0	POC1A	3smr_B	NP7_B_1000	0.453756
<i>Q15542</i>	TAF5	Q8NBT0	POC1A	3smr_B	NP7_B_1000	0.453756
<i>O75764</i>	TCEA3	P07478	PRSS2	2zzu_H	359_H_1	0.453471
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	5dbm_C	58N_C_1201	0.45293
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	5cp5_A	EB0_A_201	0.452568
<i>Q14781</i>	CBX2	Q9Y232	CDYL	4izc_B	1GZ_B_301	0.449652
<i>P17022</i>	ZNF18	Q7Z2T5	TRMT1L	4gom_D	0Y0_D_301	0.449529
<i>Q03112</i>	MECOM	Q7Z2T5	TRMT1L	4gom_D	0Y0_D_301	0.449529
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	5e3d_A	5JL_A_801	0.449284
<i>Q14781</i>	CBX2	Q9Y232	CDYL	4x3t_F	45E_F_101	0.449205
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	4tte_A	36Z_A_1204	0.449123
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	4o7e_A	2RN_A_201	0.448866
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	4uiu_A	TVU_A_1123	0.4481
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	5enj_A	5Q9_A_1501	0.446994
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	4lrg_A	1XB_A_201	0.446987
<i>Q14566</i>	MCM6	Q9UJA3	MCM8	6ut5_E	GSP_E_701	0.446427
<i>Q15542</i>	TAF5	Q9UKB1	FBXW11	6nm7_B	22L_B_201	0.445822
<i>O75764</i>	TCEA3	P07478	PRSS2	6m2n_C	3WL_C_401	0.443426
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	5rjo_A	1P8_A_1501	0.441818
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	4qsx_A	38S_A_1207	0.441227
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	5enc_A	5QD_A_1503	0.440676
<i>O76071</i>	CIAO1	Q8TED0	UTP15	5eam_B	5MN_B_401	0.439803
<i>Q15542</i>	TAF5	Q8TED0	UTP15	5eam_B	5MN_B_401	0.439803

<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	5c4q_A	BMF_A_201	0.439649
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	5ctl_A	EB9_A_201	0.439529
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	4cfl_A	8DQ_A_1169	0.439367
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	5ev9_A	5SB_A_801	0.43839
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	4qsw_A	38T_A_1207	0.438237
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	4xy9_A	43U_A_202	0.437337
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	5eni_A	5QA_A_1501	0.436794
<i>O76071</i>	CIAO1	Q8NBT0	POC1A	4ql1_B	35Q_B_401	0.436336
<i>Q15542</i>	TAF5	Q8NBT0	POC1A	4ql1_B	35Q_B_401	0.436336
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	4nyv_A	15E_A_1201	0.435352
<i>P37231</i>	PPARG	P41235	HNF4A	3omp_B	W14_B_1	0.435301
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	5c89_A	4YT_A_801	0.434409
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	4tz2_A	39R_A_1205	0.433838
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	4qr4_A	BNK_A_201	0.432843
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	5a5p_A	JTF_A_2111	0.432476
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	6v0x_A	B49_A_301	0.431374
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	4a9i_C	P9I_C_1183	0.430907
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	4o77_A	2RE_A_201	0.430242
<i>O76071</i>	CIAO1	Q86VZ2	WDR5B	5j89_B	6GX_B_201	0.429744
<i>Q09028</i>	RBBP4	Q86VZ2	WDR5B	5j89_B	6GX_B_201	0.429744
<i>Q15542</i>	TAF5	Q86VZ2	WDR5B	5j89_B	6GX_B_201	0.429744
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	4q0n_E	2XD_E_803	0.429019
<i>Q14781</i>	CBX2	Q9Y232	CDYL	4izd_B	1HE_B_301	0.42638
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	5fh8_D	5XK_D_801	0.42621
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	6v0x_A	B49_A_301	0.425674
<i>O43679</i>	LDB2	Q8IV38	ANKMY2	5uij_B	TYD_B_402	0.424763
<i>O75764</i>	TCEA3	P07478	PRSS2	1dy9_B	2ZF_B_401	0.424759
<i>P37231</i>	PPARG	P41235	HNF4A	5die_B	5CJ_B_601	0.424252
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	5dx4_A	E0C_A_201	0.424127
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	4j0s_A	1H3_A_205	0.423676
<i>Q9UIG0</i>	BAZ1B	Q9UIF9	BAZ2A	4pci_A	2NJ_A_201	0.423545
<i>P15036</i>	ETS2	Q8NB46	ANKRD52	5uij_B	TYD_B_402	0.423481
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	6fgl_A	UO1_A_1901	0.422193
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	2ydw_A	WSH_A_1183	0.422052
<i>Q9UIG0</i>	BAZ1B	Q8WWQ0	PHIP	6v1k_A	5SW_A_201	0.421968
<i>P17022</i>	ZNF18	Q8N823	ZNF611	4avv_D	GHE_D_1207	0.421803
<i>P41182</i>	BCL6	Q8N823	ZNF611	4avv_D	GHE_D_1207	0.421803
<i>Q03938</i>	ZNF90	Q8N823	ZNF611	4avv_D	GHE_D_1207	0.421803
<i>Q9UIG0</i>	BAZ1B	Q9NRL2	BAZ1A	4qsv_A	THM_A_1210	0.421534
<i>O76071</i>	CIAO1	P54198	HIRA	3smr_B	NP7_B_1000	0.420051

000257	CBX4	Q9H5I1	SUV39H2	3rjw_B	CIQ_B_2000	0.418568
Q14781	CBX2	Q9H5I1	SUV39H2	3rjw_B	CIQ_B_2000	0.418568
O76071	CIAO1	P54198	HIRA	5eam_B	5MN_B_401	0.417751
Q9UIG0	BAZ1B	Q9NRL2	BAZ1A	5fh6_A	5XM_A_801	0.41576
O60828	PQBP1	Q96QZ7	MAGI1	3o5n_E	BR0_E_1	0.415122
P46937	YAP1	Q96QZ7	MAGI1	3o5n_E	BR0_E_1	0.415122
O75764	TCEA3	P07478	PRSS2	3ufa_B	VPF_B_201	0.414907
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	4j0r_A	1H2_A_205	0.414621
Q9UIG0	BAZ1B	Q8WWQ0	PHIP	4hxo_A	1A6_A_201	0.414166
P37231	PPARG	P41235	HNF4A	5aau_B	XBR_B_1547	0.413701
Q9UIG0	BAZ1B	Q9NRL2	BAZ1A	4ts8_A	XZ8_A_1203	0.412928
P37231	PPARG	P41235	HNF4A	4iu7_B	1GM_B_601	0.412808
P37231	PPARG	Q14541	HNF4G	5hcv_B	60R_B_1001	0.412061
Q9UIG0	BAZ1B	Q9NRL2	BAZ1A	5fe9_A	5WS_A_901	0.410791
Q9UIG0	BAZ1B	Q8WWQ0	PHIP	4tz2_A	39R_A_1205	0.410777
P15036	ETS2	Q96JP0	FEM1C	5uij_B	TYD_B_402	0.409812
P15036	ETS2	Q8NB46	ANKRD52	5uim_B	T3Q_B_402	0.409715
P03372	ESR1	P22105	TNXB	3t80_C	CTN_C_401	0.408727
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	4hxr_A	1A4_A_201	0.406867
Q9UIG0	BAZ1B	Q8WWQ0	PHIP	5fdz_A	5X0_A_901	0.406476
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	4j3i_A	1K0_A_201	0.40593
Q9UIG0	BAZ1B	Q8WWQ0	PHIP	5dfc_A	EAM_A_503	0.405777
Q9UIG0	BAZ1B	Q9NRL2	BAZ1A	4c66_A	H4C_A_1168	0.405181
P37231	PPARG	P41235	HNF4A	4ivw_B	1GJ_B_601	0.404261
O76071	CIAO1	Q9Y263	PLAA	5eam_B	5MN_B_401	0.404032
Q15542	TAF5	Q9Y263	PLAA	5eam_B	5MN_B_401	0.404032
P50281	MMP14	Q8N119	MMP21	4ogk_E	THM_E_301	0.403992
O76071	CIAO1	P62873	GNB1	5eam_B	5MN_B_401	0.403387
O76071	CIAO1	Q13347	EIF3I	4x9d_E	U5P_E_104	0.403052
P63165	SUMO1	Q9UBT2	UBA2	4ajh_C	88S_C_1335	0.402817
O76071	CIAO1	Q8TED0	UTP15	4ql1_B	35Q_B_401	0.402607
Q15542	TAF5	Q8TED0	UTP15	4ql1_B	35Q_B_401	0.402607
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	5iid_A	6BK_A_801	0.402341
Q9UIG0	BAZ1B	Q9NRL2	BAZ1A	5fh7_A	5XL_A_803	0.401412
Q9UIG0	BAZ1B	Q9NRL2	BAZ1A	5a5r_A	NP8_A_2111	0.401086
P63165	SUMO1	Q9UBT2	UBA2	4ajl_D	88W_D_1333	0.400976
Q9UIG0	BAZ1B	Q9UIF9	BAZ2A	5eni_A	5QA_A_1501	0.400948

**Table 3.S3 Predicted PCIs involving PrePPI predicted master regulator binding proteins.**  
PCIs with high LT-scanner scores and low sequence similarity scores which contact at least 5 residues on the target protein and which overlap the predicted master regulator binding site.

## **Chapter 4: Structural specificity analysis of druggable pockets within protein-protein interaction interfaces**

### **4.1 Introduction**

While the structural conservation of ligand binding sites enables PrePCI to identify surface regions on new proteins which likely bind similar ligands, this conservation often complicates the development of targeted therapeutics, as drugs intended to target one protein also bind and inhibit other proteins with similar binding sites<sup>148, 149</sup>. One particularly notable example of this phenomenon occurs in the family of protein kinases, a class of approximately 500 enzymes which phosphorylate other proteins<sup>150, 151</sup>. These phosphorylation events are often key mediators of signaling pathways by which the cell senses and integrates environmental stimuli and coordinates responses ranging from cell migration to mitosis and apoptosis<sup>152, 153</sup>. Consistent with these critical roles in mediating cell survival signals, dysregulation of protein kinase activity is associated with an array of diseases including numerous oncologic, inflammatory and degenerative diseases<sup>154, 155</sup>. Pharmacologic inhibition of specific protein kinases without disrupting normal cellular networks is therefore clinically desirable. However, most kinases share a highly similar ATP-binding site and consequently, drugs intended to inhibit a specific kinase by targeting its ATP binding site often bind to and inhibit other kinases as well<sup>150, 156</sup>. This off-target activity can lead to widespread disruption of cellular signaling networks, mediating unwanted side-effects and toxicity and limiting clinical utility<sup>157</sup>. Accordingly, it is of great interest to develop targeted inhibitors which

can specifically interact with a single protein, or comparatively small subset of proteins, within a protein family.

Numerous strategies have been developed to identify unique regions within protein pockets which can confer pharmacologic selectivity. Grid-based features describing the local pharmacophoric environments within protein pockets have been used to highlight regions which are unique to the target protein and may potentially be targeted for increased drug specificity<sup>158-160</sup>. Similarly, research in volumetric descriptions of cavities<sup>161, 162</sup> and cavity electrostatic isopotentials<sup>163, 164</sup> has been applied to identify subpockets that are unique to individual proteins within a family and which may confer specificity to a specific substrate. Use of multiple protein structures, including homology models, enables the construction of binding site similarity distributions and the detection of subpockets that are significantly distinct from other proteins within the family which may indicate selectivity determining regions<sup>165, 166</sup>.

While such methods have the promise of identifying specificity determining regions within active sites when such regions exist, an emerging alternative strategy is to develop small molecules capable of modulating interactions between proteins, either disrupting or stabilizing the complexes through which they mediate their biological effects<sup>167</sup>. Such “edgetic perturbations” present the opportunity to more finely perturb individual protein functions as, rather than entirely eliminating the activity of a target protein, a PPI modulating drug may more precisely target some functionality by directly disrupting its interaction with some proteins while preserving interaction with others<sup>168, 169</sup>. Moreover, targeting PPIs may help to expand the range of druggable targets, as many proteins of clinical interest have intrinsically disordered regions which complicate identification of small molecules which will efficiently bind them<sup>170</sup>. While such proteins may be

difficult to target directly, it may be possible to alter their activity indirectly by instead targeting the proteins to which they bind.

Despite the potential benefits, disrupting PPIs with small molecules has historically been considered particularly challenging due to the distinct physicochemical features of protein-protein interfaces. PPI interfaces, often between 1500 and 3000Å<sup>2</sup>, are typically significantly larger than the 300-1000Å<sup>2</sup> surface area characteristic of drug-like small molecules<sup>171, 172</sup>. Moreover, PPI interfaces are often relatively flat and featureless compared to conventional active site pockets, lacking substantive grooves which can encapsulate a small molecule and provide sufficient stabilizing contacts to prevent diffusion away<sup>173-175</sup>. Finally, the interior surfaces of PPI interfaces tend to be predominantly hydrophobic and lacking specificity determining features such as charged residues or hydrogen bond donors/acceptors which could enable the design of selective drugs<sup>171</sup>. Further complicating the design process, whereas enzymes typically bind to endogenous metabolites which can be used as an initial lead for subsequent pharmacologic optimization, PPIs rarely have such conjugate small molecules and consequently any inhibitor would likely need to be designed from scratch<sup>176</sup>. Consequently, developing specific small molecules capable of binding a protein-protein interface stably and specifically enough to disrupt individual protein-protein interfaces has long been considered particularly challenging.

The above challenges notwithstanding however, numerous PPI inhibitors have been developed recently with several gaining clinical approval. The first approved PPI inhibitor was venetoclax which targets the pro-apoptotic Bcl-2, preventing its association with anti-apoptotic proteins and thereby inducing cell death<sup>177</sup>. Since it was first approved in 2016 for use in chronic lymphocytic leukemia, venetoclax has been successfully applied to treating additional malignancies including small lymphocytic lymphoma and acute myeloid leukemia<sup>178</sup>. Moreover,

several additional PPI inhibitors have also been approved for various cancers in the intervening years<sup>179</sup>. In addition to these clinical successes, databases such as the 2P2I<sup>180, 181</sup>, DLiP-PPI<sup>182</sup> and IPPI-DB<sup>183</sup> containing experimentally validated PPI disrupting small molecules have been created to facilitate the discovery of novel PPI inhibitors and therapeutics.

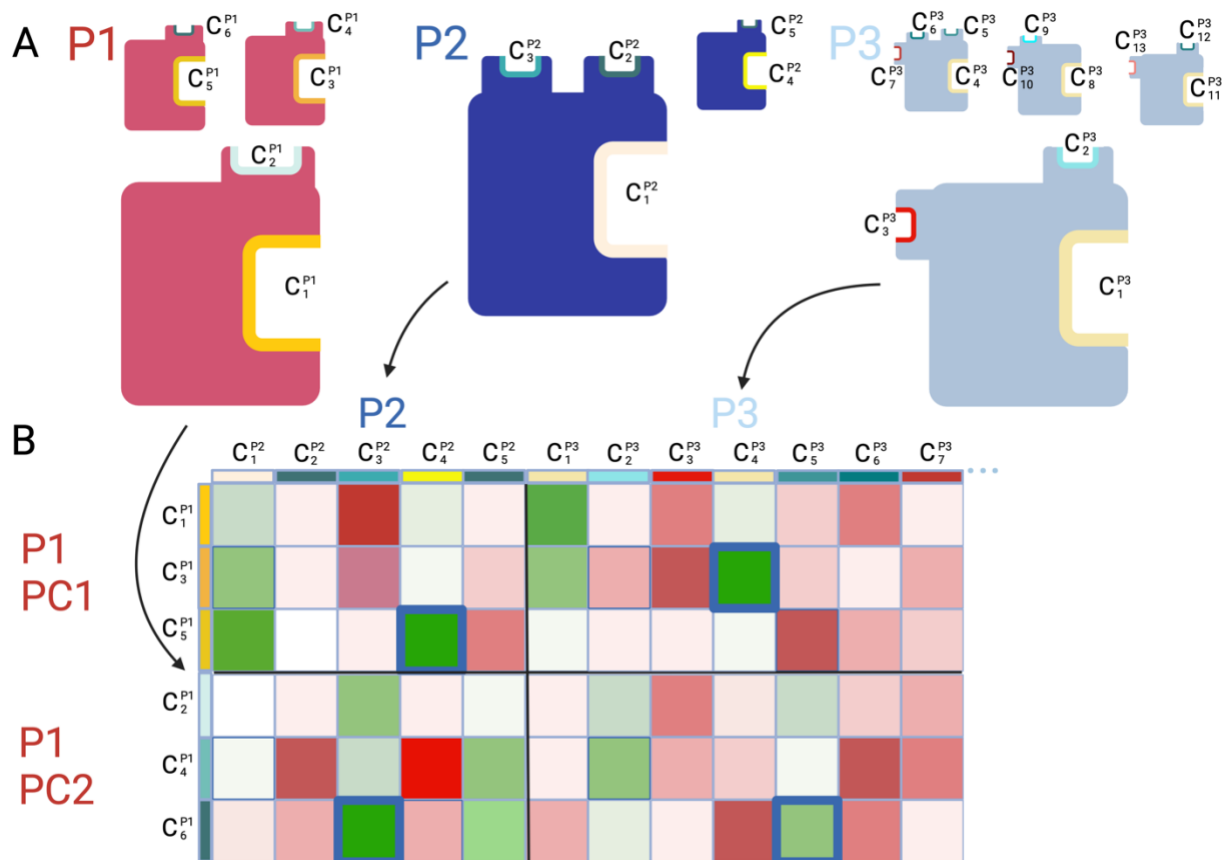
With the growing clinical and scientific successes in identifying small molecules capable of modulating PPIs, several groups have sought to analyze the prevalence and properties of PPI interfaces that are disruptable by small molecules. Systematic structural analyses of the PDB suggest that druggable pockets within PPI interfaces may be more widespread than previously recognized. Da Silva et al. identified over 160,000 pockets within and near PDB protein-protein interfaces that were predicted to be druggable<sup>95, 184</sup>. They found that these protein-protein interfacial pockets are, in general, structurally distinct from conventional active site pockets, finding only two similar pocket pairs out of 1.98 billion comparisons between conventional ligandable sites and PPI druggable pockets, prompting the authors to posit that PPI inhibitors may occupy as-of-yet unexplored regions of chemical space, orthogonal to conventional inhibitors<sup>95</sup>. Moreover, Skolnick and Zhou found that 58% (48%) of homomultimeric (heteromultimeric) interfaces contained a concave region large enough to support ligand-binding on both sides of the interface and another 14% (19%) of homomultimeric (heteromultimeric) interfaces contained a concave pocket on one side of the interface<sup>185</sup>. Moreover, the authors observed numerous instances where metabolites bound to residues associated with protein binding in PPIs involved in separate metabolic pathways, which the authors speculate could be a mechanism coupling the metabolic state of the cell to the protein-protein interactome<sup>185</sup>.

The pharmacologic successes in disrupting a selection of PPIs as well as the systematic analyses of protein structures described above support the plausibility of developing orthosteric,

PPI inhibitors on a larger scale. While such small molecules may enable the targeting of new classes of proteins, it remains uncertain whether PPI inhibitors will target these proteins any more selectively than conventional active site drugs. Some studies note that, whereas the evolution of active sites is constrained by the geometric and physicochemical requirements of catalysis, protein-protein interfaces are able to tolerate a greater degree of sequence variability. In particular, partially solvent exposed residues in the rim of the PPI interface mutate at a faster rate than core residues, increasing the structural diversity of interfaces and the plausibility of greater drug selectivity<sup>186</sup>.<sup>187</sup>. Additionally, Johnson and Karanicolas analyzed low-energy pockets generated through a Rosetta structural sampling protocol within the PPI binding interfaces of Bcl-2 family proteins and found that, while most proteins in the family contained a structurally similar binding pocket, the fluctuations in binding pocket structure for each protein were not identical, instead producing structurally distinct pockets in each protein which the authors suggest could confer selectivity<sup>188</sup>.

In contrast, numerous studies of protein-protein interfaces present evidence highlighting the conservation of protein-protein interfaces, potentially undermining the prospects of selectivity. Studies by our lab have found that protein-protein interfaces tend to be highly geometrically conserved, including between proteins from different protein families<sup>189, 190</sup>. Guharoy and Chakrabarti found that hotspot residues, or residues whose mutation to alanine causes a large change in the PPI's binding free energy, are more highly conserved than the rest of the interface<sup>191</sup>. Moreover, using a method called Multiple Structural Alignments of Protein-Proteins Interactions (MAPPIS), Shulman-Peleg et. al. found that conserved physicochemical features could be used to predict the location of such hotspot residues<sup>192</sup>. The physicochemical conservation of these sites questions the purported structural diversity of PPI pockets underlying the expectations of increased small-molecule specificity.

For this reason, we were interested in assessing whether the pharmacologically relevant pockets within protein-protein interfaces are as structurally conserved as active site pockets, or whether there is a greater degree of structural variability which might make PPI pockets more unique and confer increased pharmacologic specificity. We used the tool Volsite<sup>193, 194</sup> to generate grid-based pharmacophoric feature maps of pockets associated with typical ligand binding sites from the scPDB database<sup>195-197</sup> as well as within human PPIs from the PPIome database<sup>95</sup> (see Materials for more details). We then clustered each protein's pockets using pairwise similarity scores to identify a structurally distinct set of pockets associated with each protein. Finally, we estimated the expected number of off-target proteins that would be likely affected by pharmacologically targeting each unique pocket, both within the PFam family<sup>198</sup> of the domain associated with the pocket, as well as in all proteins outside the family as a control (Figure 4.1 and Methods). We found that PPIs pockets tended to be associated with fewer predicted off-targets than conventional active site pockets, however, we note that this result was family specific with some protein families displaying no significant difference between PPI and conventional pockets. We discuss possible sources of bias which may have impacted our results and suggest means of ameliorating these biases in future studies.



**Figure 4.1 Pocket-based comparison method for identifying off-targets.**

(A) Beginning with a protein, P1, all other proteins in the database (either scPDB or PPIome) within the same PFM family (P2, P3) are identified and druggable pockets are detected in all proteins using Volsite. Each pocket is labeled as “C”, for cavity (to reserve the label “P” for protein), where the superscript ( $P_i$ ) indicates which protein the cavity occurs in and the subscript ( $j$ ) indicates the arbitrary index labeling the pocket. (B) Pairwise similarity metrics are computed for all of P1’s pockets which are subsequently hierarchically clustered to identify structurally unique pocket clusters (PC1, PC2). For each pocket cluster beginning with PC1, all pairwise pocket similarities between the cluster pockets and pockets in P2 are computed (black outline), the maximal observed similarity is identified (dark green square with bold blue outline) and P2 is considered an off-target of the pocket if this similarity exceeds a cutoff. This process is repeated for all remaining proteins within the PFM family (ie P3) to determine the fraction of family proteins which are off-targets of the pocket cluster. This procedure is then repeated for each pocket cluster (ie PC2), and finally using all other proteins within the family (ie P2 and P3) as the reference protein to generate distributions of the fraction of off-target proteins for each unique pocket in the family. A similar procedure is performed to identify off-targets proteins from other PFM families (ie using proteins P2, P3 from different PFM families) in order to estimate the background rate of off-targets defined by this method.

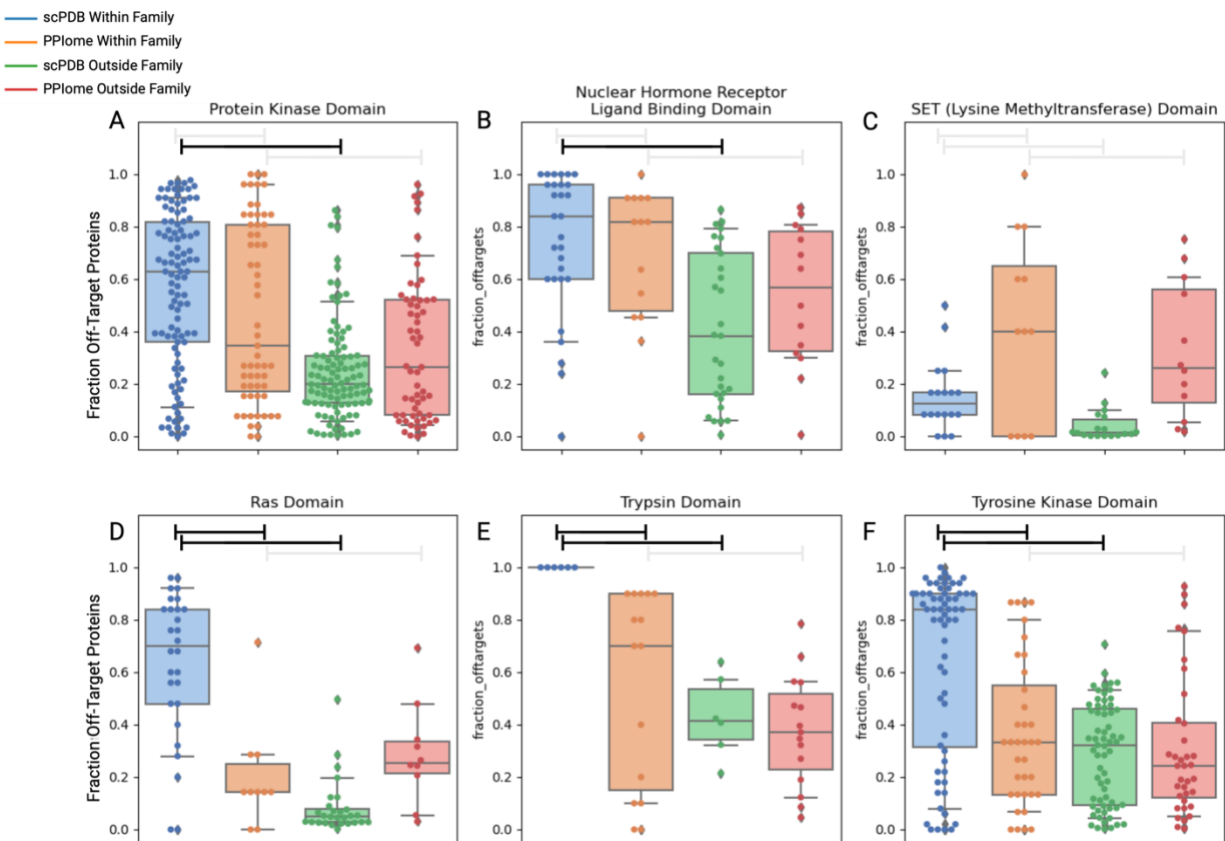
## 4.2 Results

### 4.2.1 Druggable PPI pockets are associated with fewer off-targets than active site pockets in some protein families.

To compare the expected number of off-target proteins which would be affected by pharmacologically targeting conventional active site pockets compared to pockets occurring in PPI interfaces, we first used Volsite to identify active site and PPI-associated druggable pockets in the scPDB<sup>195-197</sup> and PPIome<sup>95</sup> databases respectively. We then computed similarity metrics for all pairs of pockets for each protein and used hierarchical clustering to group the pockets into structurally distinct clusters. We subsequently compared the pockets in each cluster to pockets from all other protein domains within the same PFam family to determine what fraction of within-family domains contained at least one pocket with a similarity greater than 0.44, a threshold suggested by the Desaphy et al. to distinguish between similar and dissimilar pockets<sup>194</sup>. Proteins containing a pocket with a similarity exceeding 0.44 were considered as likely off-targets of any drug designed to target the original cavity cluster. We repeated this procedure using proteins which did not belong to the same PFam family in order to estimate the background expectation of cavity similarity as well as to assess the sensitivity of the method to discriminate true off-target sites.

Using the above approach, we estimated the fraction of proteins within a protein family, as well as outside the family, which would likely be an off-target of a drug designed to target each distinct pocket cluster, focusing on families with at least 5 proteins in both the scPDB and the PPIome. This analysis included pockets for the PFam families composed of protein kinase domains (PF00069), ras domains (PF00071), trypsin domains (PF00089), nuclear hormone receptor ligand binding domains (PF00104), SET domains (PF00856) and tyrosine kinase domains (PF07714). The distributions of within-family off-targets for scPDB pockets were significantly greater than

the distribution of out-of-family off-targets for all of the above families except SET domains (Figure 4.2, blue vs green), generally consistent with the observation that active sites are highly structurally conserved within families. In contrast, the distributions of within-family off-targets for PPIome pockets were never significantly different from the distribution of off-targets outside of the PFam family (Figure 4.2, orange vs red), suggesting that PPI pockets tend to be less tightly conserved and that the similarity PPI associated pockets within families are as similar to one another as pockets from randomly chosen pairs of proteins. However, we also found that the fraction of expected within-family off-targets for PPIome pockets from protein kinase, ras and nuclear hormone receptor ligand binding domains were not significantly different from scPDB off-targets (Table 4.S1 and Figure 4.2 A-C, blue vs orange), whereas they were significantly different in the cases of ras domains, trypsin domains and tyrosine kinase domains (Table 4.S1 and Figure 4.2 D-F, blue vs orange). These results suggest that the plausibility of developing targeted PPI inhibitors may depend on the specific target of interest. For example, it may be more feasible to specifically disrupt PPIs involving tyrosine kinases and ras kinases than nuclear hormone receptors or kinases overall, as pockets belonging to the former groups are associated with significantly fewer off-targets than the latter. Moreover, the insignificant differences between scPDB and PPIome cavities for SET domains suggest that active site inhibitors may be sufficiently unique to achieve targeted therapeutics for this class without attempting to develop PPI inhibitors.

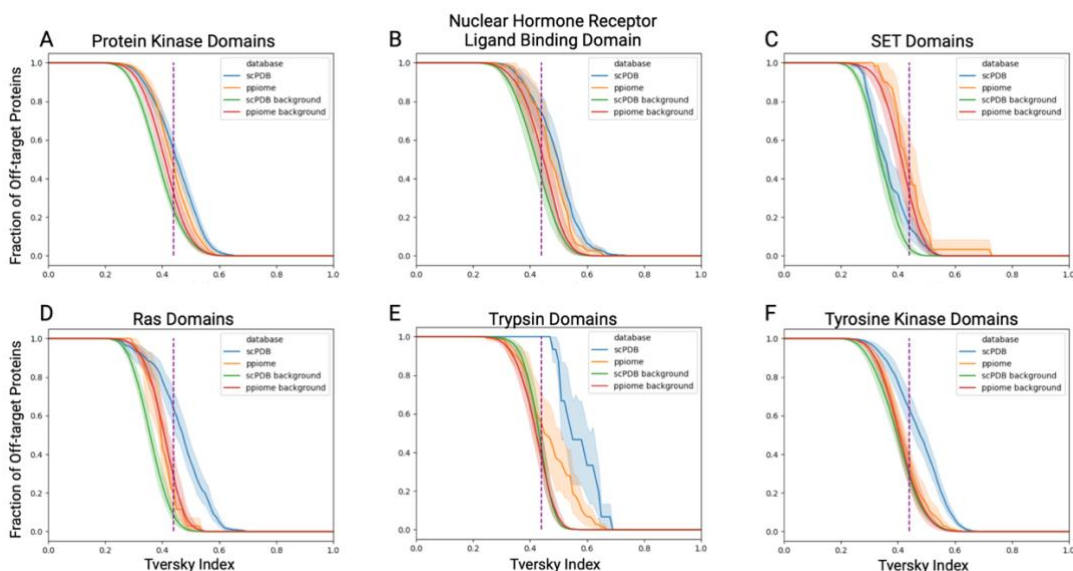


**Figure 4.2 Distribution of expected number of off-targets.**

For each unique cavity, the number of proteins within (outside of) the same PFam family containing a pocket with a Tversky Index > 0.44 was obtained and normalized by the number of proteins within (outside) the family for (A) Protein Kinases, (B) Nuclear Hormone Receptor Ligand Binding Domains, (C) SET Domains, (D) Ras Domains, (E) Trypsin Domains and (F) Tyrosine Kinase Domains. Black bars above pairs of distributions indicate that the two distributions are statistically significantly different (Bonferroni adjusted Mann-Whitney U Test p-value < 0.05). Gray bars indicate pair of distributions were not statistically significantly different.

The above analysis was based on using a threshold separating similar pockets from dissimilar pockets of 0.44. By allowing the threshold to vary from 0 to 1, we observed the expected fraction of proteins containing an off-target pocket as a function of the similarity threshold. From the resulting plots, we observe that, within the range of roughly 0.4 to 0.6, the above results appear largely independent of the threshold used. PPIome off-target curves associated with kinase and nuclear hormone receptor ligand binding domains were uniformly below the scPDB curves but

with overlapping confidence intervals, undermining the plausibility of greater drug selectivity for PPI pockets in these domains. Meanwhile PPIome off-target curves associated with SET domains were generally above the curves for scPDB pockets, reinforcing the conclusion that greater drug selectivity might be obtained by targeting the SET domain active sites compared to its PPI pockets. In contrast, ras and tyrosine kinases showed fairly robust separation between PPIome curves, which were virtually indistinguishable from background, and scPDB curves, suggesting that targeting PPIs in ras and tyrosine kinases might yield increased specificity and fewer off-targets than targeting active site pockets. Similarly, trypsins displayed a fairly string separation between scPDB and PPIome pockets, with PPIome pockets becoming nearly indistinguishable from random below a similarity threshold of 0.44.



**Figure 4.3: Fraction of off-target proteins as a function of Tversky Index.**

For each unique pocket cluster, the maximal similarity between a pocket within the cluster and a pocket in a partner protein was determined. The number of off-target proteins was plotted as a function of the Tversky Index threshold, where a protein was considered an off-target if it contained a pocket with a similarity greater than the threshold. Analyses performed for the Pfam Families (A) Protein Kinases, (B) Nuclear Hormone Receptor Ligand Binding Domains, (C) SET Domains, (D) Ras Domains, (E) Trypsin Domains and (F) Tyrosine Kinase Domains. Vertical line corresponds to Tversky Index = 0.44 to facilitate comparison with Figure 4.2.

## 4.3 Discussion

We have presented preliminary work in analyzing the structural heterogeneity of druggable pockets which occur in protein-protein interfaces. We observed a tendency of PPIome pockets being associated with fewer predicted off-targets than corresponding pockets in the scPDB. However, we note that this effect is family specific, occurring most prominently in tyrosine kinase, trypsin and ras domains while being statistically insignificant in protein kinase, SET and nuclear hormone receptor ligand binding domains.

Notably, more than half of out-of-family proteins were identified as off-targets for many pockets, especially kinases and nuclear hormone receptor ligand binding domains (Figure 4.2, green and red). While it is possible that these domain families simply contain pockets which are structurally similar to many other proteins, it is suspicious that so many off-targets are found outside of the PFam family. It is possible that a more stringent cutoff criterion for off-target could be used, such as 0.6 which nearly eliminates all predicted out-of-family off-targets in both the scPDB and PPIome for all families (Figure 4.3), however additional benchmarking would be required to identify an optimal threshold. Alternatively, the current analysis is based on a similarity calculation which includes an adjustable parameter set to 0.95 (see Methods). While this parameter value was found to be optimal when originally studied in 2012<sup>194</sup>, it is possible that a value which better separates on-targets from off-targets for updated datasets could be identified. Finally, alternative cavity comparison methods<sup>199-205</sup> may provide more robust distinction between on-target and off-target cavities.

### 4.3.1 Possible Sources of Bias

The above results based on available PDB structures suggest that, for some protein families, it may be possible to design drugs targeting PPI interfaces which would result in fewer

off-targets than drugs directed at conventional active sites. However, this analysis is complicated by at least three potential sources of bias, sampling bias, holo-structure bias and methodological bias, which we describe in more detail here and suggest how future work might address these biases in chapter 6.

#### **4.3.1.1 Sampling Bias**

Some proteins in the PDB have been crystallized numerous times, with each structure presenting the opportunity for slight structural distortions in pocket geometry which increase the probability of yielding a structurally unique pocket. Proteins which are less frequently crystallized lack these additional opportunities, thus limiting the number of distinct pockets which can be found in these proteins relative to their more popular homologs. Consequently, when searching for proteins with potential off-targets pockets, underrepresented proteins will have sampled less of the pocket structural space that is available to them, decreasing the probability of finding a highly similar pocket. Since this would tend to underestimate the number of off-targets, this sampling bias could potentially shift the off-target distributions in Figure 4.2 (Figure 4.3) downward (leftward), giving an erroneously favorable estimation of the selectivity of PPI binding sites. While this effect is, in principle possible for the scPDB as well, we expect that it would be less pronounced than for PPIome due to the holo-structure bias presented in the next section.

#### **4.3.1.2 Holo-structure Bias**

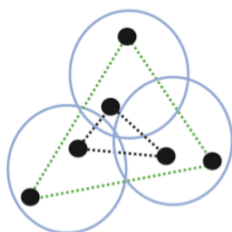
The scPDB is a database derived from PCI holo-structures. As described in Section 3.1.2, proteins can undergo large conformational changes upon ligand binding which result in surface geometries which are distinct from the unbound state<sup>58</sup>. Moreover, binding of a ligand to a protein surface results in the formation of specific interactions which can restrict the conformational space of the protein. Pockets bound by ligands may therefore be less structurally variable than the

corresponding sites in the absence of ligand and, consequently, scPDB pockets may be artificially more similar to one another due to the presence of a stabilizing ligand than apo-structure pockets occurring of the same sites. As such ligand-bound structures are not generally available for PPI pockets, the observed increase in the number of off-targets associated with scPDB pockets compared to PPIome pockets may be due in part to the fact that the former were detected in holo-structures while the latter were detected in apo-structures. Repeating the analysis using pockets at ligand-binding sites derived from apo-structures may provide a fairer comparison between conventional active site pockets and PPI pockets.

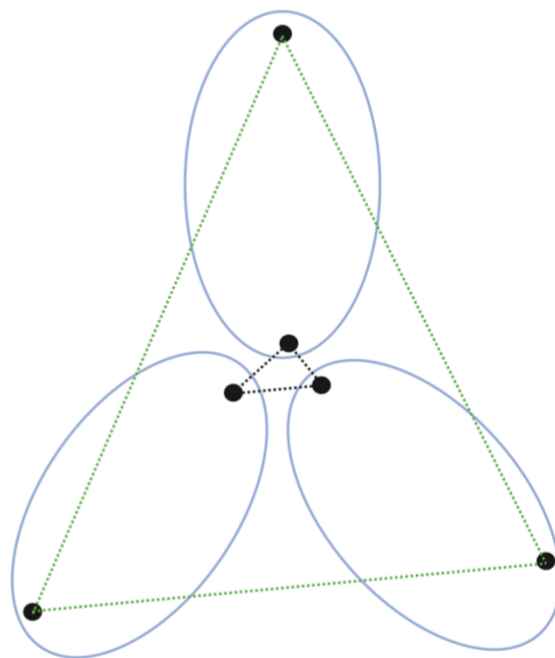
#### **4.3.1.3 Methodological bias**

In identifying whether a protein contained a pocket which exceeded a given threshold, we searched for off-targets by identifying the pocket in a potential off-target protein which was most similar to a pocket from the cluster. While this approach provides a worst-case estimate of the number of off-target proteins which would likely result from targeting a given cavity, it neglects Johnson and Karanicolas's observation that proteins within the Bcl-2 family, despite sharing a structurally similar pocket, also display fluctuations which can enable the formation of structurally distinct pockets which could impart drug selectivity. As a result, in identifying the most similar pocket between two sets, we implicitly ignored the possibility developing inhibitors that specifically target the least similar pockets. As shown in Figure 4.4, if PPIome pockets exhibit larger structural variations around a shared common core pocket, it may be possible to identify more structurally distinct cavities than the scPDB, even if the minimum distance between clusters remains unchanged. Possible methods for ameliorating this bias are suggested in Chapter 6.

A



B



#### Figure 4.4 Hypothetical Methodological Bias.

The analysis used in this study identified the most similar pockets between two proteins without taking pocket fluctuations which might be unique to individual proteins into account. (A) Hypothetical schematic of the range of possible active-site pockets for three proteins in a two-dimensional abstract pocket space. The shortest distance between pockets (black dashed lines) is of similar magnitude as the largest distances (green dashed lines) since active site fluctuations are fairly small. (B) Hypothetical schematic of the range of possible PPI pockets for three proteins in two-dimension abstract pocket space. Here, the shortest distance between pockets is identical to that of active sites, (black dashed lines) however the fluctuations are comparatively larger, resulting in larger maximal distances between pockets (green dashed lines) and unique pockets in each protein that are ignored if minimal distance (i.e. maximal similarity) is the primary selection criteria.

## 4.4 Materials and Methods

### 4.4.1 Materials

*Volsite/Shaper* - Volsite is a grid-based method for identifying and assessing the druggability of pockets on proteins surfaces<sup>193, 194</sup>. It begins by instantiating a grid around a protein of interest and then identifying points exterior to the protein surface using ray-casting. Points exterior to the

protein surface are then evaluated for buriedness, where 120 uniformly distributed rays are emanated from the point and the number which impinge on the protein surface are tallied. Buried points are those where the number of rays which impinge on the protein exceeds a cutoff, 40 by default<sup>194</sup>. Cavities are then defined as densely connected regions of buried grid points exterior to the protein surface. Each point of the cavity is subsequently labeled as a pharmacophoric feature which is the inverse of the nearest protein atom to the grid point. For example, points near a hydrogen bond donor would be labeled as a hydrogen bond acceptor while points neighboring a positive charge would be labeled as a negative charge. In this way, the cavities defined by Volsite describe the spatial orientation of complementary pharmacophoric features that a drug targeting that site would need to have<sup>194</sup>.

Volsite cavities were compared using the Shaper tool which describes the pocket's pharmacophoric features using a smooth Gaussian function from OpenEye's ShapeTK package and attempts to maximize the overlap between two pockets<sup>194</sup>. Once two pockets are aligned, they can be compared using a Tversky Index defined as:

$$TI = \frac{I_{AB}}{I_{AB} + \alpha D_A + \beta D_B}$$

where  $I_{AB}$  are the number of points the two cavities have in common,  $D_A$  and  $D_B$  are the number of points distinct to cavities A and B respectively, and  $\alpha$  and  $\beta$  are adjustable weight terms which sum to one and have the effect of prioritizing one cavity ( $\alpha$ ) as the reference and the other ( $\beta$ ) as the comparison. Following Desaphy et al, we used the Shaper default values of  $\alpha=0.95$  and  $\beta=0.05$  and chose the smaller of the two pockets as the reference pocket and the larger as the comparison. Previous benchmarking of Volsite and Shaper found that these tools can efficiently differentiate between similar and dissimilar pockets, and the authors recommended the use of 0.44 as a threshold for distinguishing between similar and dissimilar pockets<sup>194</sup>.

*PPIome* –The PPIome is a database of pockets within protein-protein interfaces that are predicted to be druggable<sup>95</sup>. Da Silva et al defined 4 types of cavities in their analysis: Interfacial, Rim, Allosteric and Orthosteric. Orthosteric pockets were identified by superimposing monomeric chains onto identical proteins in protein complexes. The monomeric chains were then evaluated for druggable pockets using the program Volsite and cavities which contained at least one grid point within 4.5Å of the interface were retained. In contrast, Interfacial, Rim and Allosteric cavities were identified in the complex structures and could accordingly be inhibitors or stabilizers depending on the extent of PPI interface occlusion<sup>95</sup>. We chose to focus on orthosteric pockets as possible targets of orthosteric PPI inhibitors. Summary descriptors of each cavity in the PPIome were obtained from:

[http://bioinfo-pharma.u-strasbg.fr/labwebsite/downloads/PPIome/Cavities\\_O\\_RAW.csv](http://bioinfo-pharma.u-strasbg.fr/labwebsite/downloads/PPIome/Cavities_O_RAW.csv).

Pockets corresponding to human proteins were programmatically downloaded directly from the PPIome website using curl. However, when we attempted to map the individual cavities to the protein residues they contact, we found zero neighboring residues within 4Å of cavity grid points for several pockets. Careful inspection of several examples suggested that the discrepancy typically occurred in cases of homomultimers, where it appeared that a different chain than the one identified by the PPIome was aligned to the PPI structure, however the pocket was detected in the chain described by the PPIome summary data. We chose to reperform the pocket detection as described below to ensure that all pockets considered in this study met our criteria to be considered orthosteric.

*scPDB* - The scPDB (screening PDB) is a minimally redundant database containing structures of proteins in complex with drug-like small molecules<sup>195-197</sup>. It contains formatted mol2 files corresponding to the protein and ligand structures which are directly usable by Volsite. The

scPDB database was obtained from <http://bioinfo-pharma.u-strasbg.fr/scPDB/ressources/2016/scPDB.tar.gz>. A summary table was obtained by querying the scPDB online database (<http://bioinfo-pharma.u-strasbg.fr/scPDB/FORM>) with an empty search string and exporting the resulting entries to a csv-formatted file.

*PDBFam* - The PDBFam<sup>206</sup> database maps individual protein chains in the PDB to their respective PFam<sup>198</sup> families with residue level resolution. We used the PDBFam to prevent comparisons between pockets occurring on different domains of multidomain proteins by mapping pocket contact residues (residues with an atom within 4Å of a pocket grid point) to their associated PFam domains<sup>198, 206</sup>. PDBFam was accessed on 21 August 2023 from <http://dunbrack2.fccc.edu/ProtCiD/PDBfam/Download.aspx>.

## 4.4.2 Methods

### Cavity Detection, assignment to PFam domains and similarity calculation

For each pocket identified within a human protein in the PPIome database, the monomer associated with the pocket was aligned to its corresponding PPI chain using *ska*<sup>97</sup>, protonated using the Protoss API<sup>207, 208</sup> and converted to mol2 files with *chimera*<sup>209</sup>. Pockets were then detected using *Volsite*<sup>193, 194</sup> in protein mode using the protein mol2 as an argument as follows:

```
ichem volsite <protein.mol2>
```

For each pocket, any residue with at least one atom within 4Å of a cavity grid point was considered a contacting residue and orthosteric pockets were defined as any pocket with a volume between 230Å<sup>3</sup> and 1350Å<sup>3</sup> (following Desaphy et al.)<sup>194</sup> which contacted at least 3 residues of the target protein and 2 residues on the partner protein.

For the scPDB, pockets corresponding to each ligand binding site were identified using Volsite in ligand mode by including the ligand mol2 as an argument as follows:

```
ichem volsite <protein.mol2> <ligand.mol2>
```

and as for the PPIome, contacting residues were defined as any residue with at least one atom within 4Å of a pocket grid point.

Cavity contact residues obtained during the cavity detection described above were cross-referenced with the domain definitions in PDBFam<sup>206</sup>. Each cavity was assigned to any domain which contained at least one contact residue. PFam families which contained at least 5 proteins were retained for further analysis.

We used the Shaper<sup>194</sup> tool to compute Tversky Indices for all pairwise combination of pockets in human proteins within each database, using the smaller of the two pockets as the reference and the larger as the comparison.

### Clustering Protein Pockets

To reduce the redundancy and sampling bias associated with different proteins being represented by varying numbers of PDB files, we clustered all of the pockets for each protein in the scPDB and PPIome databases to identify groups of structurally unique pockets. Clustering was performed separately for each database. For each protein, all pairwise Tversky Indices were computed using Shaper as described above. Tversky Indices were converted to distances by subtracting the observed Tversky Index from 1:

$$d(P1, P2) = 1 - TI(P1, P2)$$

and aggregated into a single distance matrix. We performed average-linkage hierarchical clustering using a threshold of 0.65 which corresponds to a Tversky Index of 0.35 which was recommended as a more conservative threshold for identifying dissimilar pockets<sup>194</sup>.

### Estimation of the number of off-target proteins for each pocket

For each pocket cluster, we identified the most similar pocket on every other protein within the same PFam family. Proteins which contained a pocket with a Tversky Index  $> 0.44$  with one of the cluster pockets were considered to be an off-target of that cavity. We repeated the analysis with proteins outside of the PFam family to determine the expected rate of “off-target” proteins obtained via this analysis among proteins we do not expect to be similar. The number of off-targets was then normalized by the number of proteins in each group to obtain the fraction of off-target proteins in each set. We repeated this process for every protein in each PFam family with at least 5 proteins in both the scPDB and the PPIome, generating distributions of the fraction of off-target proteins for each pocket, both within and outside of the protein family. We compared the distribution of scPDB off-targets to the distribution of PPIome off-targets, as well as comparing the scPDB and PPIome distributions to their respective out-of-family background distributions, using a Mann-Whitney U Test (Figure 4.2). P-values were adjusted using a Bonferroni correction to account for multiple hypothesis testing (Table 4.S1). We also performed threshold independent visualization by varying the threshold from 0 to 1 and displaying the fraction of each protein family which is predicted to be an off-target at any intermediate threshold (Figure 4.3).

## 4.5 - Supplemental Information

**A**

<i>Off-target Distribution Comparisons</i>	<i>Protein Kinase Domain</i>	<i>Ras Domain</i>	<i>Trypsin Domain</i>	<i>Nuclear Hormone Receptor Ligand Binding Domain</i>	<i>SET (Lysine Demethylase) Domain</i>	<i>Tyrosine Kinase Domain</i>
<i>scPDB/PPIome</i>	0.066	<b>2.27E-04</b>	<b>2.06E-04</b>	0.137	0.064	<b>3.85E-05</b>
<i>scPDB/scPDB Random</i>	<b>3.16E-13</b>	<b>4.33E-08</b>	<b>1.39E-03</b>	<b>3.85E-05</b>	6.29E-03	<b>1.36E-08</b>
<i>PPIome/PPIome Random</i>	0.012	0.091	0.062	0.038	0.375	0.182

**B**

<i>Off-target Distribution Comparisons</i>	<i>Protein Kinase Domain</i>	<i>Ras Domain</i>	<i>Trypsin Domain</i>	<i>Nuclear Hormone Receptor Ligand Binding Domain</i>	<i>SET (Lysine Demethylase) Domain</i>	<i>Tyrosine Kinase Domain</i>
<i>scPDB/PPIome</i>	1	<b>4.08E-03</b>	<b>3.70E-03</b>	1	1	<b>6.93E-04</b>
<i>scPDB/scPDB Random</i>	<b>5.69E-12</b>	<b>7.79E-07</b>	<b>0.025</b>	<b>6.92E-04</b>	0.113	<b>2.45E-07</b>
<i>PPIome/PPIome Random</i>	0.208	1	1	0.687	1	1

**Table 4.S1 Statistical significance of off-target distributions.**

(A) Raw p-values obtained by Mann-Whitney U test comparing the distribution of off-target proteins (maximal Tverskey Index > 0.44) for each pocket cluster. (B) Bonferroni adjusted p-values computed as minimum(1, 18 \* (raw p-value)). Adjusted p-values < 0.05 in bold.

# **Chapter 5: Estimation of volatile odorant Binding Affinities to wildtype and mutant MhOR5 using Free Energy Perturbation**

## **5.1 Introduction**

### **5.1.1 Olfactory Receptors**

Chemical space is vast and discriminating between distinct, yet highly similar, small molecules can be essential for survival. Accordingly, accurate detection of specific small molecules via olfaction poses a unique challenge as, unlike photoreceptors which detect light waves that vary in frequency and amplitude but are otherwise qualitatively similar, olfactory receptors (ORs) must be able to recognize a diverse array of odorants which are qualitatively distinct from one another<sup>210-212</sup>. This discriminatory capacity the olfactory system is typically achieved by encoding the identity of each small molecule as the specific combination of ORs activated by the odorant. Rather than identifying a single compound via highly specific interactions, ORs tend to recognize common structural and physicochemical features shared by numerous small molecules. Similarly, each odorant does not bind to a single olfactory receptor, rather to an entire repertoire, with each OR identifying a distinct set of features<sup>213</sup>. Though two small molecules may bind and activate an overlapping subset of ORs, the entire repertoire of ORs activated by a given odorant is unique to that odorant. This strategy enables highly specific odorant discrimination using a finite collection of ORs as the identity of the odorant can be inferred from the simultaneous activation of individually promiscuous olfactory receptors<sup>210-212</sup>.

### 5.1.2 Biotechnological chemosensors

There has been recent interest in co-opting the discriminatory capacity of the olfactory system in order to develop devices to detect and identify specific small molecules. In contrast to vertebrate ORs which are G-protein coupled receptors<sup>214, 215</sup>, insects typically utilize heterotetrameric ion channels<sup>216, 217</sup> composed of a highly divergent OR subunit responsible for detecting the ligand, and a highly conserved Olfactory Receptor Coreceptor (orco) subunit<sup>218, 219</sup>. This combination of a highly conserved functional subunit with a highly divergent detection subunit facilitates the evolution of new olfactory receptor chains by decoupling of ligand sensing, which requires a diverse repertoire of odorant binding sites, from channel opening, which depends on preserved functional activity of the channel<sup>219</sup>.

Curiously, the olfactory receptor of the jumping bristletail, *Machilis hrabei*, is a homotetrameric ion channel and does not contain an orco subunit, likely due to the divergence of *M. hrabei* from the neopteran insect lineage prior to the evolution of orco<sup>213</sup>. The structure of the broadly tuned *M. hrabei* OR5 (MhOR5) was recently solved by Dr. Vanessa Ruta in the unliganded apo states, as well as in complex with the ligands eugenol and DEET (PDB files 7LIC, 7LID and 7LIG respectively). The Ruta lab also performed GCaMP fluorescence dose-response assays measuring channel activity in response to a diverse panel of odorants, as well as measuring the effect of MhOR5 odorant binding site mutations on channel activity in response to DEET and eugenol exposure. From these assays, they obtained the half maximal effective concentration (EC<sub>50</sub>) characterizing the concentration of ligand which produced half the maximal fluorescence signal, providing an estimate of binding affinity and a quantitative benchmark for retrospective Free Energy Perturbation analysis<sup>213</sup>.

Here, we describe an ongoing effort to study the ligand binding preferences of MhOR5 using FEP. Specifically, we are interested in evaluating whether FEP+ can be used to characterize the binding affinity of an array of infection associated small volatile odorants to MhOR5, with the long-term interest of accelerating the development of biotechnological platforms for detecting specific small molecules.

### **5.1.3 Prospective Ligand FEP**

Previously, Dr. Pierre Devlaminck from Dr. Richard Friesner's lab performed extensive Relative Binding FEP (RBFEP) simulations of the MhOR5 system in complex with a selection of the ligands studied by the Ruta group. The goal of these studies was to assess the reliability of FEP to reproduce the GCaMP activity data described above. Of the 55 compounds tested by Dr. Ruta, Dr. Devlaminck chose 33 to analyze by FEP, splitting them into 6 congeneric series of small molecules. These series were comprised of phenols, benzoyls, aliphatic alcohols, aliphatic carbonyls, aliphatic esters and heterocycles. Eugenol was used as the reference ligand for benzoyl and phenol compounds. For each of the remaining 4 congeneric series, a representative ligand was chosen and IFD-MD was used to produce an initial pose which would form the basis for subsequent RBFEP calculations. For compounds for which the EC<sub>50</sub> was measurable within the dynamic range of the experiment, these analyses were fairly successful, achieving an R<sup>2</sup> of 0.55 for the correlation of predicted change in binding affinity ( $\Delta\Delta G$ ) and observed  $\Delta\log(\text{EC}_{50})$ .

Here, we describe ongoing efforts to prospectively predict the binding affinity of a new set of ligands with MhOR5. 155 small molecule ligands were provided by the Bill and Melinda Gates Foundation, comprising a set of small volatile odorants detected in patients infected with Malaria, Tuberculosis and Sars-CoV-2. These ligands spanned a diverse array of chemotypes, which we visually clustered into the following classes: alkanes/alkenes, aliphatic alcohols, aliphatic

carbonyls, esters and ethers, carboxylates, thioethers, cyclohexane derivatives, cyclohexene derivatives, phenyl derivatives (combining the above benzoyl and phenol series into a single group), naphthalene-like systems, purine-like systems and camphene-like systems. Of these groups, we chose to focus on alcohols, esters, aliphatic carbonyls and phenyl derivatives for which sufficient retrospective validation data was available to perform RBFEP and assess the reliability of results. Beginning with either 7LID for phenyl derivatives or Dr. Devlaminck's IFD-MD structures for the remaining series, we used Dr. Devlaminck's optimized membrane generation and relaxation protocol (described further in Methods) to embed the ligand-bound, tetrameric MhOR5 channel in a membrane and solvate the entire system. From this initial structure, we subsequently performed RBFEP for the new prospective ligands, including the retrospective ligands in the calculation to calibrate the  $\Delta G$  predictions and to serve as an internal control.

### **5.1.4 Retrospective Protein FEP**

In addition to the above ligand FEP experiments, Dr. Devlaminck previously evaluated whether FEP+ could reproduce the effect of several binding site mutations on MhOR5 channel activity in response to DEET and eugenol exposure. Predicted binding affinities of mutant MhOR5 channels with DEET were generally consistent with the observed activity measurements, however for eugenol there were several mutations, most notably involving residues M209 and I213, which produced severe and persistent outliers. Although the mutations generally reduced binding affinity (increased  $EC_{50}$ ), FEP+ typically predicted these mutations increased binding affinity (negative  $\Delta\Delta G$ ). Numerous strategies were attempted to optimize the simulation and improve the correlation with experiment, including adding positional restraints and using ligand FEP to interconvert between DEET and eugenol in the background of mutant receptors, however the outliers persisted.

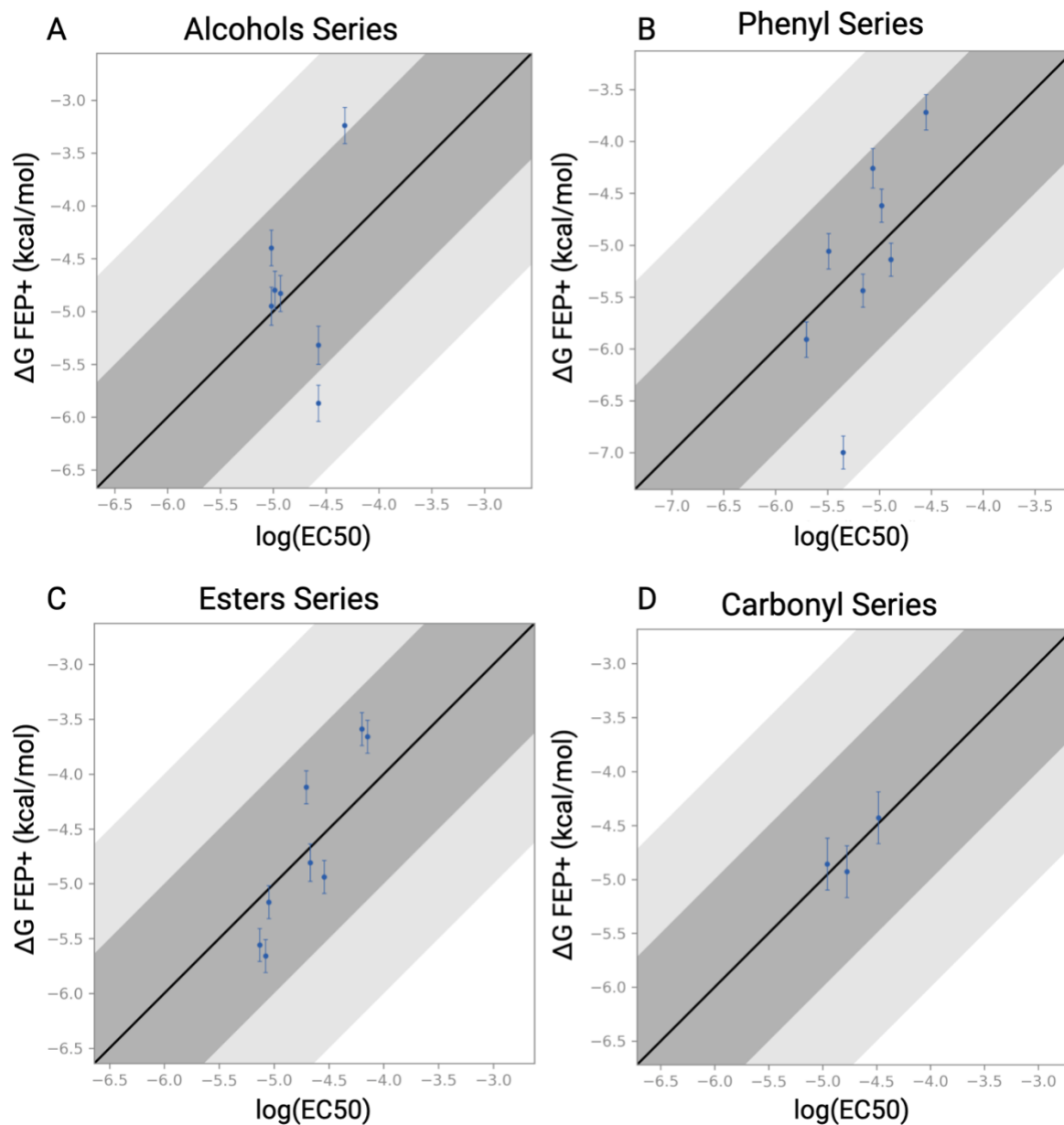
Previously, Ioannidis et al. investigated the effect of membrane lipid composition on FEP+ affinity predictions for a series of small aminoadamantane derivatives with the Influenza A M2 proton channel. They found that the correlation coefficient between the experimental affinities and FEP+ predictions were highly dependent on the specific lipid used, degrading from 0.88 when using a DMPC membrane to 0.516 when a DPPC membrane was used instead<sup>131</sup>. We therefore reasoned that the use of a POPC membrane might not be ideal for this system and that other membrane lipids might improve the correlation of FEP predictions with experiment. To that end, we reran protein mutation FEP for mutations involving the residues M209 and I213 using all four membrane lipids compatible with FEP+: 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC), 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine (POPE), 1,2-dipalmitoyl-sn-glycero-3-phosphocholine (DPPC) and 1,2-dimyristoyl-sn-glycerol-3-phosphocholine, (DMPC).

## **5.2 Results**

### **5.2.1 Ligand FEP**

We used FEP+ to estimate the binding free energies for 84 of 155 disease associated volatile odorants, focusing on those which were structurally similar to four congeneric series previously studied by Dr. Devlaminck: aliphatic alcohols, aliphatic carbonyls, esters and phenyl derivatives (Figure 5.1, Table 5.1 and Table 5.2). Because the chirality of several of these 84 compounds was unspecified, we used all stereoisomers of each compound whose SMILES string representation was ambiguous, resulting in 90 compounds in total, 32 of which were associated with retrospective activity data while the remaining 58 constituted prospective affinity predictions. Consistent with Dr. Devlaminck's previous studies, we used the compounds 1-octanol, 2-heptanone, ethyl hexanoate and eugenol as reference compounds for the alcohols, carbonyls, esters

and phenyl derivatives respectively. For each series,  $\Delta G$  estimates for retrospective and prospective ligands were obtained by fitting FEP+  $\Delta\Delta G$  estimates to the GCaMP dose-response activity  $\log(\text{EC}_{50})$  measurements of retrospective compounds whose affinity with MhOR5 was measurable within the range of the assay, i.e. excluding top-of-the-assay compounds. The accuracy of  $\Delta G$  calculations for top-of-the-assay compounds was evaluated using this regression and reported in Table 5.1.



**Figure 5.1 Retrospective RBFEP estimation of ligand binding affinity to MhOR5 and benchmarking against experimental affinities.**

$\Delta G$  predicted by FEP+ versus the  $\log(\text{EC}_{50})$  obtained from GCaMP dose-response activity measurements of MhOR5 for (A) alcohols, (B) phenyl derivatives, (C) esters and (D) carbonyls. Only compounds whose affinity with MhOR5 was measurable within the dynamic range of the assay were used to fit  $\Delta G$  estimates from RBFEP  $\Delta\Delta G$  calculations. See Table 5.1 for evaluation of top-of-the-assay compounds.

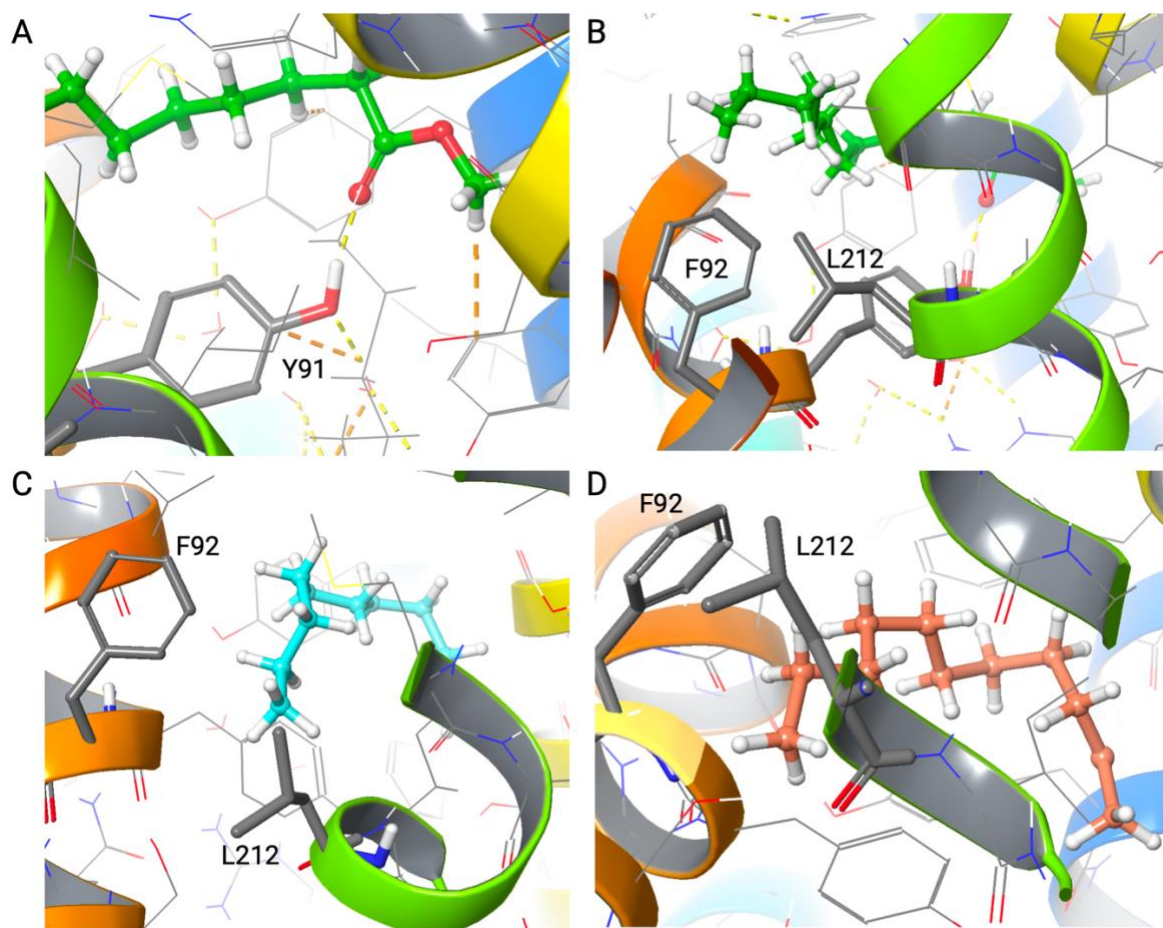
	LIGAND	FEP+ ΔG (KCAL/MOL)	PRED. ERROR	LOG(EC <sub>50</sub> )	STRUCTURE
<b>BENZOYLS/PHENOLS</b>	Eugenol	-5.91	0.17	-5.7	<chem>C=CCc1cc(OC)c(O)cc1</chem>
<b>REFERENCE: EUGENOL (7LID)</b>	acetophenone	-5.06	0.17	-5.49	<chem>CC(=O)c1ccccc1</chem>
	methyl benzoate	-7	0.16	-5.35	<chem>COC(=O)c1ccccc1</chem>
	benzaldehyde	-4.62	0.16	-4.98	<chem>O=Cc1ccccc1</chem>
	o-cresol	-3.72	0.17	-4.55	<chem>Cc1c(O)cccc1</chem>
	4-methoxyphenyl acetone	-5.14	0.16	-4.89	<chem>CC(=O)Cc1ccc(cc1)OC</chem>
	2-ethylphenol	-4.26	0.19	-5.06	<chem>CCc1c(O)cccc1</chem>
	4-ethylphenol	-5.44	0.16	-5.16	<chem>CCc1ccc(O)cc1</chem>
<b>ALCOHOLS</b>	1-octanol	-4.81	0.18	-4.99	<chem>OCCCCCCC</chem>
<b>REFERENCE: 1-OCTANOL (IFD-MD)</b>	1-hexanol	-3.25	0.17	-4.38	<chem>OCCCCC</chem>
	(R)-1-octen-3-ol	-4.84	0.17	-4.93	<chem>C=C[C@H](O)CCCCC</chem>
	1-pentanol	-2.15	0.24	N/A	<chem>CCCCCO</chem>
	trans-3-hexen-1-ol	-2.24	0.22	N/A	<chem>OCC/C=C/CC</chem>
	(R)-3-octanol	-4.41	0.17	-5.02	<chem>CC[C@@H](O)CCCCC</chem>
	(S)-3-octanol	-4.96	0.18	-5.02	<chem>CC[C@H](O)CCCCC</chem>
	(R)-linalool	-5.88	0.17	-4.57	<chem>CC(C)=CCC[C@@](C)(O)C=C</chem>
	(S)-linalool	-5.33	0.18	-4.57	<chem>CC(C)=CCC[C@](C)(O)C=C</chem>
<b>CARBONYLS</b>	2-heptanone	-4.86	0.24	-4.96	<chem>CC(=O)CCCCC</chem>
<b>REFERENCE: 2-HEPTANONE (IFD-MD)</b>	2-undecanone	-6.24	0.29	N/A	<chem>CC(=O)CCCCCCCCC</chem>
	decanal	-6.56	0.32	N/A	<chem>O=CCCCCCCCC</chem>
	hexanal	-4.43	0.24	-4.48	<chem>O=CCCCCC</chem>
	sulcatone	-4.93	0.24	-4.77	<chem>CC(C)=CCCC(=O)C</chem>
<b>ESTERS</b>	ethyl hexanoate	-5.56	0.15	-5.13	<chem>CCOC(=O)CCCCCC</chem>
<b>REFERENCE: ETHYL HEXANOATE (IFD-MD)</b>	butyl acetate	-5.17	0.15	-5.05	<chem>CC(=O)OCCCC</chem>
	ethyl acetate	-2.16	0.15	N/A	<chem>CC(=O)OCC</chem>
	ethyl butyrate	-3.66	0.15	-4.15	<chem>CCOC(=O)CCC</chem>
	isobutyl acetate	-4.12	0.15	-4.71	<chem>CC(=O)OCC(C)C</chem>
	isopropyl tiglate	-4.81	0.17	-4.67	<chem>C\C=C(C)\C(=O)OC(C)C</chem>
	methyl hexanoate	-5.66	0.15	-5.08	<chem>COC(=O)CCCCCC</chem>
	propyl acetate	-3.59	0.15	-4.2	<chem>CC(=O)OCCC</chem>
	prenyl acetate	-4.94	0.15	-4.54	<chem>CC(C)=CCOC(=O)C</chem>
	methyl laurate	-7	0.22	N/A	<chem>COC(=O)CCCCCCCCCCC</chem>

**Table 5.1 FEP+  $\Delta G$  predictions for retrospective ligand series**

$\Delta G$  predicted by FEP+ and the  $\log(EC_{50})$  obtained from channel GCaMP activity measurements of MhOR5 for alcohols, phenyl derivatives, carbonyls and esters. Only compounds whose affinity with MhOR5 was measurable within the dynamic range of the assay were used to fit  $\Delta G$  estimates from RBFEP  $\Delta\Delta G$  calculations.  $\log(EC_{50})$  for compounds lacking measurable affinities (i.e. top-of-the-assay compounds) denoted as N/A.

All series demonstrated reliable estimation of binding free energies for compounds with measurable affinities, achieving mean absolute errors of 0.6, 0.1, 0.6 and 0.4 kcal/mol for the phenyl, carbonyl, alcohol and ester series respectively (Figure 5.1, Table 5.1). The discrepancy between experimental  $\log(EC_{50})$  and the  $\Delta G$  predicted by FEP+ was within 1kcal/mol for nearly all compounds, though notable outliers include 1-hexanol and (R)-linalool from the alcohol series falling just outside 1kcal/mol error and methyl benzoate from the phenyl derivatives with an error of 1.65 kcal/mol.

Evaluation of top-of-the-assay compounds revealed that the predicted affinity of methyl laurate from the ester series and decanal and 2-undecanone from the carbonyl series deviated significantly from the experimental  $EC_{50}$  (Table 5.1). FEP+ predicts these compounds bind to MhOR5 with fairly high affinity (-6 to -7kcal/mol) despite undetectable MhOR5 activity in GCaMP experiments. Similar inaccuracies for these ligands had been previously observed by Dr. Devlaminck. Consistent with his findings, analysis of the trajectories associated with these small molecules revealed increased hydrophobic contacts for methyl laurate, decanal and 2-undecanone. Methyl laurate appeared to adopt a fairly stable pose, forming a hydrogen bond with Y91's side-chain hydroxyl group while its aliphatic tail appeared to embed itself within a hydrophobic crevice between two transmembrane helices and forming stable hydrophobic contacts with residues F92 and L212 (Figure 5.2 A-B). Similarly, decanal and 2-undecanone formed stable hydrophobic contacts with residues F92 and L212 (Figure 5.3 C-D), and adjacent hydrophobic residues.



**Figure 5.2 Representative poses of methyl laurate, decanal and 2-undecanone bound to MhOR5.**

(A) Hydrogen bond formed between methyl laurate and Y91 (B) Aliphatic tail of methyl laurate contacting F92 and L212. (C) Aliphatic tail of decanal contacting F92 and L212. (D) Aliphatic tail of 2-undecanone embedded between two helices below (orange and green helices) and F92 and L212 above.

Prospective predictions for novel odorants were generally consistent with our understanding of the hydrophobic and broadly tuned nature of the MhOR5 binding pocket. Hydrophilic compounds like methanol and propanol were predicted to bind poorly to the MhOR5 channel whereas bulkier and more hydrophobic compounds, like ethenylbenzene, were predicted to bind more tightly (Table 5.2). This prompted us to evaluate the correlation of the logP with the experimental affinity measurements where available, as well as  $\Delta G$  predicted by FEP+ (Figure

5.3). The  $\log(\text{EC}_{50})$  of retrospective compounds measurable within the dynamic range of the GCaMP assay yielded low correlation with their respective  $\log P$  ( $R^2 = 0.07$ ) whereas the correlation of  $\log(\text{EC}_{50})$  with the predicted  $\Delta G$  for prospective compounds ( $R^2 = 0.68$ ) and both retrospective and prospective compounds aggregated together ( $R^2 = 0.64$ ) was substantially higher. This dramatic increase in correlation is likely driven by the increased range of predicted affinities relative to experimentally observed affinities, reflecting the expectation that high hydrophobicity is a necessary condition for binding MhOR5 but insufficient to distinguish compounds that bind with high affinity from similarly hydrophobic weak binders. The compounds with the greatest deviation between the  $\Delta G$  predicted by FEP+ and the  $\Delta G$  inferred from the  $\log P$  regression were both enantiomers of 2-butyl-1-octanol (Figure 5.3 B-C). Dose-response measurements and structural analyses of compounds whose affinity predicted by FEP+ deviates significantly from a regression based on  $\log P$ , like 2-butyl-1-octanol, may provide additional insights into the structural features of molecular recognition by MhOR5 beyond hydrophobicity. Experimental measurement of binding affinities to evaluate our prospective predictions will be performed by the Ruta group.

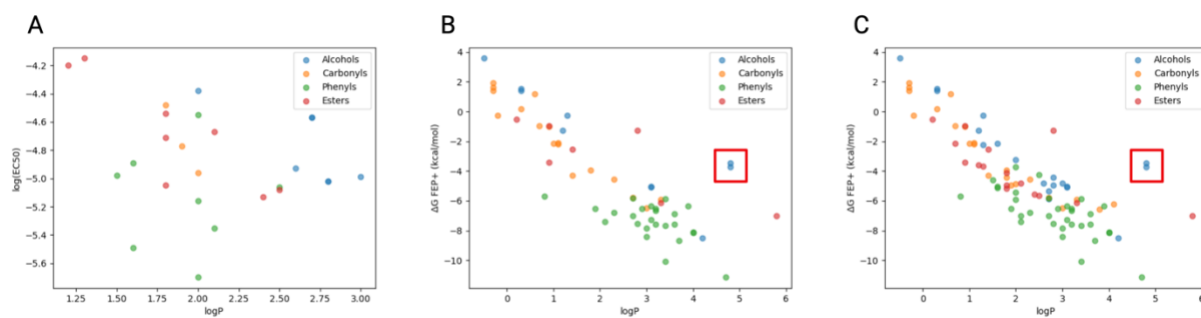
	CAS	LIGAND	FEP+ $\Delta G$	PRED. ERROR	STRUCTURE
<b>PHENYL DERIVATES REFERENCE: EUGENOL (7LID)</b>	108-88-3	toluene	-5.84	0.18	<chem>Cc1ccccc1</chem>
	100-41-4	ethylbenzene	-7.27	0.22	<chem>CCc1ccccc1</chem>
	103-65-1	propylbenzene	-8.66	0.2	<chem>CCCc1ccccc1</chem>
	100-42-5	styrene	-6.52	0.18	<chem>C=Cc1ccccc1</chem>
	673-32-5	1-phenyl-1-propyne	-7.86	0.21	<chem>CC#Cc1ccccc1</chem>
	95-47-6	o-xylene	-6.34	0.22	<chem>Cc1c(C)cccc1</chem>
	611-15-4	2-methylstyrene	-6.53	0.19	<chem>C=Cc1c(C)cccc1</chem>
	108-38-3	m-xylene	-6.66	0.21	<chem>Cc1cc(C)ccc1</chem>
	106-42-3	p-xylene	-7.6	0.23	<chem>Cc1ccc(C)cc1</chem>
	620-14-4	3-ethyltoluene	-7.57	0.19	<chem>CCc1cc(C)ccc1</chem>
	535-77-3	m-cymene	-8.16	0.18	<chem>CC(C)c1cc(C)ccc1</chem>
	64268-28-6	1-(2,2-Dimethylpropyl)-4-ethenylbenzene	-11.15	0.19	<chem>CC(C)(C)Cc(cc1)ccc1C=C</chem>

	108-67-8	mesitylene	-5.85	0.25	<chem>Cc1cc(C)cc(C)c1</chem>
	526-73-8	1,2,3-trimethylbenzene	-6.9	0.22	<chem>Cc1c(C)cccc1C</chem>
	95-63-6	1,2,4-trimethylbenzene	-8.41	0.27	<chem>Cc1c(C)ccc(C)c1</chem>
	27831-13-6	4-ethenyl-1,2-dimethylbenzene	-7.66	0.2	<chem>C=Cc1cc(C)c(C)cc1</chem>
	25619-60-7	1,2,3,4-tetramethylbenzene	-8.09	0.2	<chem>Cc1c(C)c(C)ccc1C</chem>
	104-93-8	4-methylanisole	-7.02	0.23	<chem>Cc1ccc(cc1)OC</chem>
	152477-96-8	2,5-dimethylbenzaldehyde-2,4-DNPH	-7.42	0.29	<chem>O=Cc1c(C)ccc(C)c1</chem>
	73513-56-1	1,4-dichlorobenzene	-10.07	0.18	<chem>Clc1ccc(Cl)cc1</chem>
	121-98-2	methyl p-anisate	-6.78	0.18	<chem>COC(=O)c1ccc(cc1)OC</chem>
	93-60-7	methyl nicotinate	-5.7	0.19	<chem>COC(=O)c1ccncc1</chem>
	80-46-6	4-tert-amylphenol	-6.35	0.18	<chem>CCC(C)(C)c1ccc(O)cc1</chem>
	94-30-4	ethyl para-anisate	-7.55	0.18	<chem>CCOC(=O)c1ccc(cc1)OC</chem>
	134-20-3	methyl 2-aminobenzoate	-6.54	0.2	<chem>COC(=O)c1c(N)cccc1</chem>
<b>ALCOHOLS</b>	67-56-1	methanol	3.58	0.26	<chem>CO</chem>
<b>REFERENCE: 1-OCTANOL (IFD-MD)</b>	71-23-8	1-propanol	1.53	0.27	<chem>CCCO</chem>
	75-84-3	2,2-Dimethyl 1-propanol	-0.24	0.21	<chem>CC(C)(C)CO</chem>
	3913-02-8	(R)-2-butyl-1-octanol	-3.74	0.24	<chem>CCCC[C@H](CO)CCCCC</chem>
	3913-02-8	(S)-2-butyl-1-octanol	-3.45	0.32	<chem>CCCC[C@H](CO)CCCCC</chem>
	629-06-1	1-chloroheptane	-8.49	0.22	<chem>CCCCCCCCl</chem>
	104-76-7	(S)-2-ethyl-1-hexanol	-5.02	0.22	<chem>CC[C@H](CO)CCCC</chem>
	104-76-7	(R)-2-ethyl-1-hexanol	-5.09	0.21	<chem>CC[C@@H](CO)CCCC</chem>
	67-63-0	2-Propanol	1.41	0.28	<chem>CC(C)O</chem>
	584-02-01	3-Pentanol	-1.28	0.22	<chem>CCC(O)CC</chem>
<b>CARBONYLS</b>	75-07-0	Acetaldehyde	1.4	0.43	<chem>O=CC</chem>
<b>REFERENCE: 2-HEPTANONE (IFD-MD)</b>	123-38-6	Propanal	1.2	0.31	<chem>O=CCC</chem>
	123-72-8	Butyraldehyde	-0.99	0.26	<chem>O=CCCC</chem>
	66-25-1	hexanal	-3.96	0.29	<chem>O=CCCCCC</chem>
	124-13-0	Octanal	-5.78	0.25	<chem>O=CCCCCCCC</chem>
	124-19-6	nonanal	-5.92	0.24	<chem>O=CCCCCCCCC</chem>
	78-93-3	2-Butanone	0.16	0.26	<chem>CC(=O)CC</chem>
	78-85-3	Methacrolein	-0.94	0.25	<chem>C=C(C=O)C</chem>
	513-86-0	(S)-3-hydroxy-2-butanone	1.64	0.25	<chem>CC(=O)[C@H](C)O</chem>
	513-86-0	(R)-3-hydroxy-2-butanone	1.94	0.25	<chem>CC(=O)[C@@H](C)O</chem>
	96-17-3	(S)-2-methylbutanal	-2.11	0.29	<chem>O=C[C@H](C)CC</chem>

	96-17-3	(R)-2-methylbutanal	-2.17	0.3	O=C[C@H](C)CC
	87994-87-4	3-methylbutanal	-2.16	0.25	O=CCC(C)C
	623-36-9	methylpent-2-enal	-4.28	0.26	O=C/C(C)=C/CC
	123-42-2	4-Hydroxy-4-methylpentan-2-one	-0.27	0.31	CC(=O)CC(C)(C)O
	68937-52-0	Citral	-6.5	0.3	O=C\C=C(C)\CCC=C(C)C
	111-71-7	Heptanal	-4.55	0.29	O=CCCCCCC
<b>ESTERS/ETHERS</b>	79-20-9	methyl acetate	-0.54	0.19	CC(=O)OC
<b>REFERENCE: ETHYL HEXANOATE (IFD-MD)</b>	108-21-4	isopropyl acetate	-3.41	0.18	CC(=O)OC(C)C
	80-62-6	methyl methacrylate	-2.54	0.25	C=C(C)C(=O)OC
	188595-68-8	tert-butyl-methyl ether	-0.95	0.19	CC(C)(C)OC
	4744-11-0	1,1-dipropoxypropane	-1.27	0.17	CCCOC(CC)OCCC
	2639-63-6	hexyl butyrate	-6.15	0.28	CCCC(=O)OCCCCC

**Table 5.2 FEP+  $\Delta G$  predictions for prospective ligand series**

FEP+ predicted  $\Delta G$  of binding to wildtype MhOR5 for alcohols, phenyl derivatives, carbonyls and esters. Only compounds whose affinity with MhOR5 was measurable within the dynamic range of the assay were used to fit  $\Delta G$  estimates from RBFEP  $\Delta\Delta G$  calculations.  $\log(\text{EC}_{50})$  for compounds lacking measurable affinities (ie top-of-the-assay compounds) denoted as N/A.



**Figure 5.3 Comparison of compound logP and experimental/predicted binding affinity.**

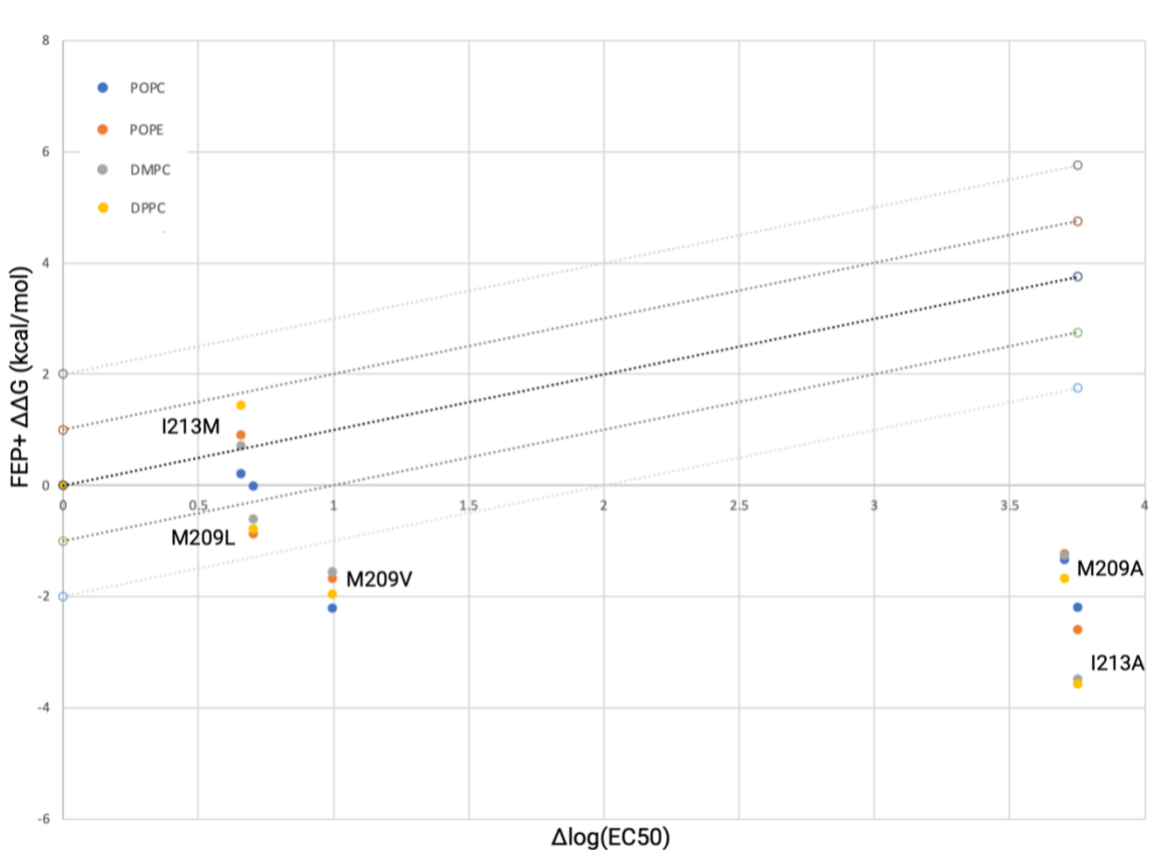
(A) Experimental  $\log(\text{EC}_{50})$  versus the  $\log P$  of each retrospective compound within the alcohol, phenyl, carbonyl and ester series. (B)  $\Delta G$  predicted by FEP+ versus the  $\log P$  of each prospective compound within the alcohol, phenyl, carbonyl and ester series. (C)  $\Delta G$  predicted by FEP+ versus the  $\log P$  of both retrospective and prospective compounds from the alcohol, phenyl, carbonyl and ester series aggregated together. Red box in B and C corresponds to 2-butyl-1-octanol.

## 5.2.2 Protein FEP

In contrast to the ligand FEP results, the protein FEP results were less reliably correlated with experimental measurements. I213M slightly weakens eugenol binding to MhOR5 and was successfully estimated by FEP+ with less than 1.0 kcal/mol error using all of four of the available lipids. In contrast, the effect of the M209L mutation which weakens eugenol binding affinity slightly more than I213M, was only predicted within 1.0 kcal/mol error using POPC. In contrast, the remaining lipids yielded predictions with negative  $\Delta\Delta G$ s (Figure 5.4). Finally, M209V yielded the largest decrease in binding affinity that was still measurable within the dynamic range of the assay and was poorly predicted in all lipid contexts;  $\Delta\Delta G$ s of roughly -2 kcal/mol were predicted for all lipids despite the experimentally observed positive  $\Delta\log(EC_{50})$ . Meanwhile, both top of the assay mutations (M209A and I213A) were predicted to stabilize eugenol binding to MhOR5 with predicted  $\Delta\Delta G$ s between -1 and -4 kcal/mol despite abrogating channel activity entirely.

Analysis of the trajectories did not reveal clear reasons that might underlie the persistent prediction inaccuracies. Convergence was generally considered “Good” and REST sampling between different lambda windows was considered “Fair” by Maestro’s automated trajectory analysis tools in all membrane systems and mutations. The apo- binding sites of the wildtype and each of the mutants were generally free of water throughout the trajectories, with one or two water molecules transiently entering the binding pocket before exiting a few frames later. Trajectories associated with M209A apo-structures in POPC and DPPC membranes, as well as I213A apo-structures in POPC and DMPC, did display more persistent binding site occupation with water molecules with at least 5 water molecules occupying the binding site for several consecutive frames. However, the persistence of the prediction inaccuracy relative to experiment, including in

trajectories where the binding pocket did not flood, suggests that there may be additional sources of error influencing the prediction.

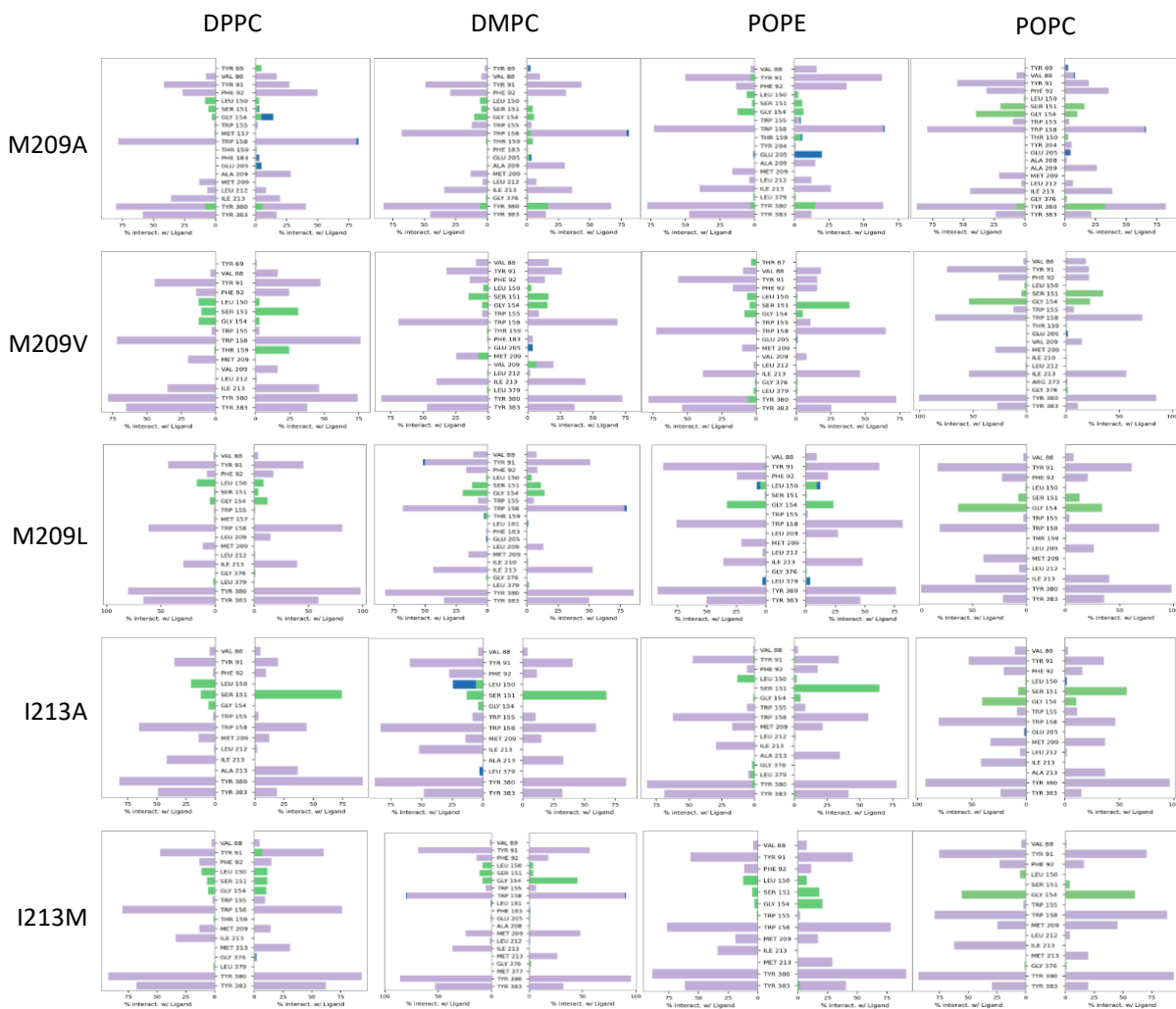


**Figure 5.4 Effect of membrane lipid composition on the ability of FEP+ to predict the effect of MhOR5 mutation on eugenol binding.**

$\Delta\Delta G$  predicted by FEP+ vs the  $\Delta\log(\text{EC}_{50})$  obtained from channel GCaMP activity measurements of mutant MhOR5 in response to eugenol exposure using each of the four lipid membranes supported by FEP+, POPC (blue), POPE (orange), DMPC<sup>81</sup> and DPPC (yellow). Activity was unmeasurable for M209A and I213A and placed at -2kcal/mol as top of the assay (I213A slightly offset to the right relative to M209A for clarity).

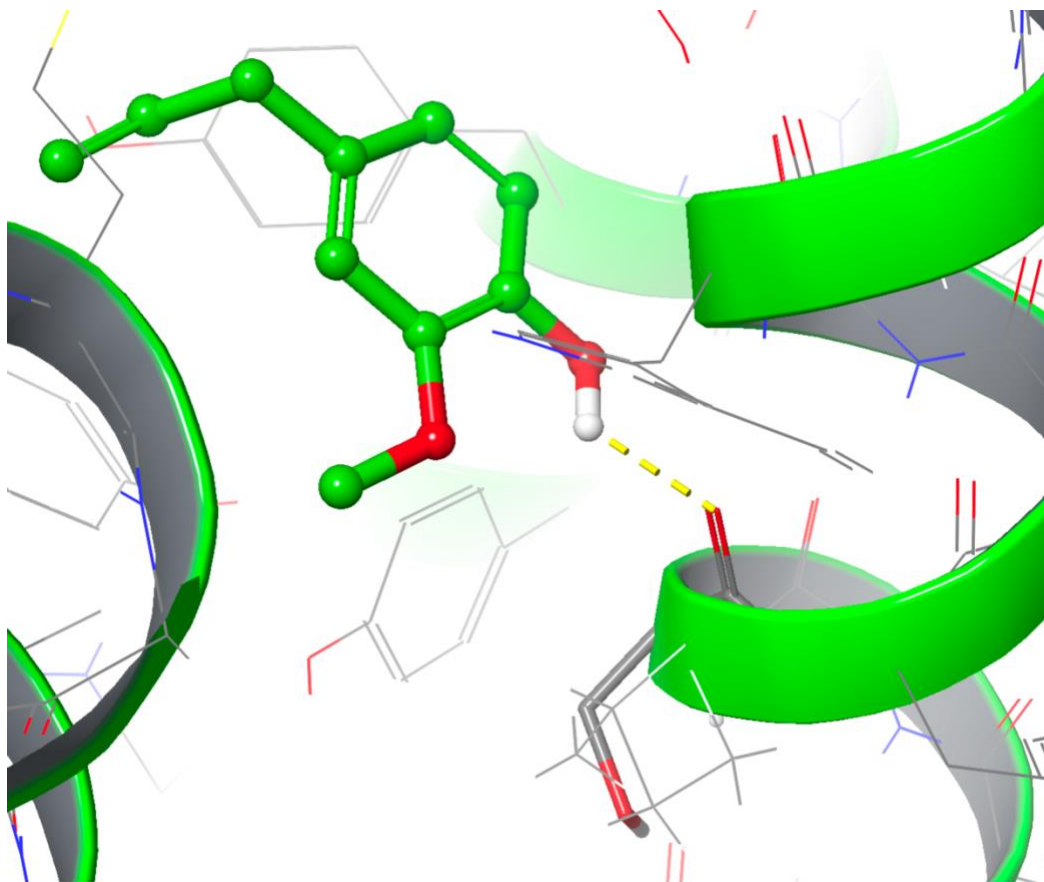
### Speculation on the source of prediction errors

Over the course of the 25ns trajectories, the distribution of contacts made between eugenol and the wildtype MhOR5 versus the I213A mutant were generally fairly similar in all four membrane systems (Figure 5.5, row 4). Hydrophobic contacts between eugenol and I213 in the wildtype were almost entirely replaced by hydrophobic contacts between eugenol and the mutant I213A residue. The notable exception to this consistency was an increase in hydrogen bonds between eugenol and S151 from roughly 0-25% of the wildtype trajectory to 50-75% of the I213A mutant trajectories (Figure 5.5, row 4). Interestingly, a similar increase in hydrogen bonding between eugenol and the backbone carbonyl of S151 occurred in the I213M simulation in the POPC membrane which was predicted to be the most stabilizing of the four simulations (Figure 5.4). Visualization of the hydrogen bond revealed that it was made between eugenol's terminal hydroxyl group and the backbone carbonyl oxygen of S151, an interaction which appeared to induce a slight distortion in the helix as the backbone carbonyl of S151 oriented towards eugenol's hydroxyl group (Figure 5.6). It is possible that this is an artifact of OPLS forcefield and use of an alternative forcefield such as Amber or CHARMM might alleviate these distortions, however these forcefields are not currently available within FEP+.



**Figure 5.5 Distribution of WT (left) vs mutant (right) MhOR5 contacts with eugenol for M209 and I213 Mutants.**

Horizontal bars indicate the fraction of frames in which each residue contacts the ligands and the type of interaction formed. Purple bars indicate hydrophobic contacts, green bars indicate hydrogen bonds and blue bars indicate water bridges.



**Figure 5.6 Representative pose of eugenol hydrogen bonding to S151 backbone carbonyl in I213A mutant MhOR5.**

I213A mutation increases the frequency of eugenol hydrogen bonding to S151's backbone carbonyl, possibly contributing to the erroneous negative  $\Delta\Delta G$  predicted by FEP+.

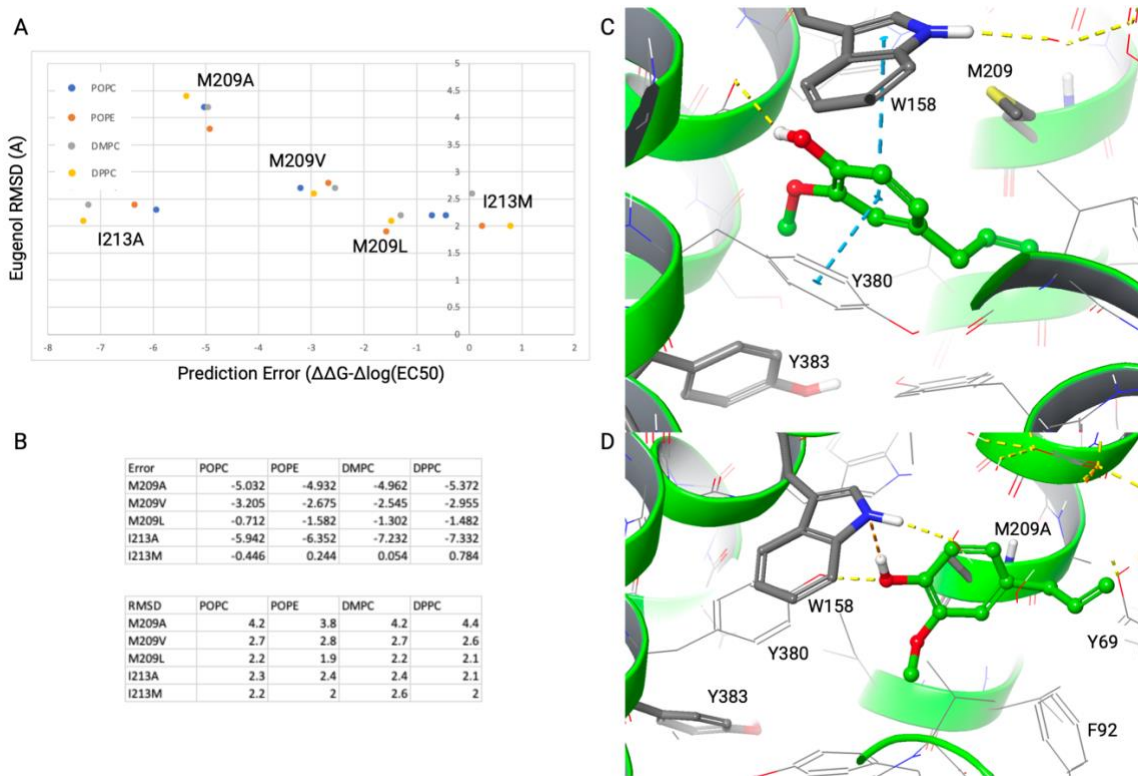
In contrast, the most substantial change observed for the M209 mutations involved hydrophobic contacts with residue Y383 (Figure 5.5, rows 1-3), which appeared to monotonically decrease as the size of the substituted amino acid decreased; Y383 formed hydrophobic contacts with eugenol most often in M209L simulations (Figure 5.5, row 3), and least often in M209A simulations (Figure 5.5, row 1), and contacted eugenol and with intermediate frequency in the M209V mutant simulations (Figure 5.5, row 2).

Additionally, although the RMSD of the ligand was fairly constant over various wildtype simulations (eugenol RMSD $\approx$ 2.2Å), we noted that the ligand RMSDs associated with M209A

trajectories were substantially greater than those from M209L simulations. We compared eugenol's RMSD in each mutant simulation against the error in the FEP+  $\Delta\Delta G$  prediction and found that greater RMSDs were associated with greater errors for M209 mutants (Figure 5.7A, B). Visualization of the trajectories suggest that the ligand is much less constrained in the M209 mutants compared to the wildtype. In particular, eugenol adopts a fairly stable pose forming pi-pi interactions with aromatic residues like Y380, Y383 and W158 which line the binding pocket (Figure 5.7 C). However, in trajectories associated with M209A, eugenol appeared to drift away from the center of the binding site, consequently breaking the pi-pi interactions with Y380, Y383 and W158, and moved towards a crevice lined by the residues Y69 and F92 (Figure 5.7 D). It is possible that the presence of wildtype M209 creates a steric barrier that prevents eugenol from accessing this crevice whereas the mutant M209A leaves open a pathway out of the binding site, allowing eugenol to embed its hydrophobic tail near residues Y69 and F92. These new hydrophobic contacts with Y69 and F92 that are accessible to eugenol in the M209A mutants could be enthalpically favorable and driving the negative  $\Delta\Delta G$  prediction. Moreover, it may be that eugenol's greater mobility within the M209A binding pocket is entropically favorable relative to the wildtype and M209L mutations where the ligand is more sterically constrained, further contributing to the negative  $\Delta\Delta G$  prediction.

Alternatively, because the dose-response experiments measure the activity of the channel rather than directly measuring binding, it is possible that eugenol is able to bind the mutant MhOR5 without activating the channel. Del Marmol et al. noted that both the background fluorescence signal was substantially increased in I213A and M209A relative to the wildtype, which they noted implied stabilization of the closed channel state (see Figure S9e and related discussion in "Architecture of the odorant binding site" section of the main text of del Marmol et al. 2021)<sup>213</sup>.

Additionally, the dynamic range of the assay appeared dramatically compressed in I213A and M209A (see Figure S9b,c of del Marmol et al. 2021)<sup>213</sup>. If sustained interaction with Y383 is required for channel opening, the decrease in eugenol contacting Y383 could result in increased binding site occupation, ie negative  $\Delta\Delta G$ , with a concomitant decrease in channel activity as observed in the GCaMP experiments.



### Figure 5.7 RMSD and trajectory analysis of MhOR5 M209 mutants.

(A) RMSD of eugenol molecule relative to initial pose versus the error in the FEP+ prediction relative to the observed  $\log(\text{EC}_{50})$  (defined as  $\Delta\Delta G - \Delta\log(\text{EC}_{50})$ ). Individual datapoints correspond to individual trajectories with different membrane lipids, with POPC in blue, POPE in orange, DMPC in gray and DPPC in yellow. Clusters of datapoints corresponding to individual mutants labeled by mutation. Eugenol mutations corresponding to M209 mutations are negatively correlated with prediction error whereas RMSD and prediction error are uncorrelated for I213 mutations. (B) Raw RMSD and prediction error data used to create the scatter plot in (A). (C) Representative eugenol pose in wildtype MhOR5 where the molecule is contained to the binding site lined by Y380, Y383 and W158. (D) Representative eugenol pose in M209A mutant MhOR5 simulation shows eugenol drifting away from the center of the binding pocket, losing pi-pi interactions with Y380, Y383 and W158, and instead embedding its hydrophobic tail outside the binding pocket and forming hydrophobic contacts with Y69 and F92.

### 5.3 Discussion

We have presented replicate retrospective evaluations of FEP+'s ability to predict the binding affinities of a panel of small odorants with wildtype MhOR5. We have also reported prospective binding affinity predictions of previously untested volatile odorants with MhOR5. These results indicate the FEP+ is accurately able to predict the effect of small perturbations in chemical structure on the binding of small molecules to MhOR5. Prospective predictions are generally consistent with expectations that small hydrophilic are unlikely to bind the MhOR5 while bulkier hydrophobic are generally predicted to bind with high affinity. Notably, FEP+ predictions fail for compounds with long aliphatic tails (ie  $\geq 10$  carbons) like methyl laurate, decanal and 2-undecanone. The accuracy of these prospective predictions will be experimentally evaluated by the Ruta group in the near future using similar GCaMP dose-response activity experiments.

In contrast, FEP+ is still unable to predict the effect of mutating residues M209 and I213 on eugenol binding affinity with MhOR5. The use of alternative lipids did not appreciably improve the reliability of the prediction. The source of the error remains unclear, however it is possible that the mutant adopts a structurally distinct conformation relative to the wildtype which are insufficiently sampled during the simulation. Obtaining additional experimental structures of the M209 and I213 mutant MhOR5, both as apo-structures and in complex with eugenol, could enable determination of whether a major conformational change occurs and provide additional structural insights into ligand binding to these mutant channels which may not be efficiently sampled by FEP. Moreover, higher resolution structures which identify bound lipids could enable more rigorous preparation of the membrane and possibly improve the reliability of FEP predictions. We have discussed obtaining these additional structures with the Ruta group and efforts to obtain them are ongoing.

## 5.4 Materials and Methods

### Initial Protein Preparation

The 7LID crystal structure containing MhOR5 in complex with the small molecule eugenol was downloaded directly from the PDB and prepared using Maestro's Protein Preparation workflow. Preprocessing steps were modified to cap termini and to fill in missing side chains using Schrodinger's Prime software, with otherwise default settings. For the reference small molecules 1-octanol, 2-heptanone and ethyl hexanoate, we used the same top-ranked IFD-MD pose used by Dr. Devlaminck for each molecule. Both the prepared 7LID and IFD-MD poses were further prepared using the membrane generation and relaxation protocol described below.

### Ligand Preparation

Ligand names and Chemical Abstracts Service (CAS) ID numbers for volatile odorants associated with Malaria, Tuberculosis and COVID-19 infection obtained in previous association studies were provided by the Bill and Melinda Gates Foundation. SMILES strings for most small molecules were obtained using the CIRPY python package which programmatically interconverts between registry IDs and molecular descriptors; SMILES strings for compounds which could not be converted with CIRPY were manually recovered from the Pubchem Chemical Database<sup>100, 101</sup>. All ligands were subsequently prepared using Maestro's Ligprep, generating all combinations of chiral centers where stereochemistry was unspecified in the SMILES string but otherwise default settings. LogP values for each compound obtained manually from Pubchem.

## Membrane Generation and Relaxation

Each initial structure (prepared 7LID for eugenol or top-ranked IFD-MD pose for other reference ligands) was imported into Maestro. For IFD-MD structures, the protein chains and IFD-MD ligand were then merged into a single object, excluding the waters and lipids generated by IFD-MD and producing a tetrameric protein system with chain A containing the reference ligand and chains B, C and D containing the original eugenol molecules from 7LID. The resulting structure was then prepared using the Desmond System Builder tool with TIP4PEW water molecules occupying a 10Åx10Åx10Å orthorhombic padding box and 0.15M NaCl added to the system. A POPC membrane was built using the “Place on Prealigned Structure” option for all ligand FEP simulations whereas the lipid was varied to assess the impact of all lipid systems (POPC, POPE, DMPC, DPPC) for protein FEP simulations. The default Desmond membrane generation protocol typically positions one or two lipid molecules within the central pore of the channel. To remove these inappropriately placed lipids, we verified the numerical ID of each lipid in the channel and deleted atoms corresponding to these lipids using python scripts and Schrodinger’s Python API.

We subsequently relaxed the resulting membrane system using the membrane relaxation protocol optimized by Dr. Pierre Devlaminck, which is the default Desmond protocol with the following modifications. 5.0 kcal/mol/Å<sup>2</sup> harmonic restraints were added to helix  $\alpha$ -carbons and ligand heavy atoms for all simulation stages. Simulation time for stage 4, corresponding to a 100K membrane equilibration step with restrained membrane z-axis motion, was increased from 100ps to 10ns. The simulation time of stage 6 corresponding to an NVT production simulation at 300K was increased from 50ps to 25ns and finally, the final 300K NP $\gamma$ T simulation of stage 8 was lengthened to 100ns.

## Ligand RBFEP

The resulting relaxed membrane system was subsequently converted to a pose viewer object compatible with FEP+ using the command:

```
$SCHRODINGER/run membrane_cms2fep.py <relaxed-membrane-system.cms> -  
ligand "chain.name Z"
```

for each of the docked ligands. For eugenol, the above command was modified to:

```
$SCHRODINGER/run membrane_cms2fep.py <relaxed-membrane-system.cms> -  
ligand "chain.name A AND residue.num 900"
```

We subsequently loaded this pose viewer object into Maestro, prepared the additional ligands of each series using Ligprep (see above) and flexibly aligned the small molecules to the reference ligand using Maestro's Ligand Alignment tool. Where possible, ligands were aligned by aligning the maximum common substructure, however if the resulting alignment was inadequate, such as by positioning key groups (ie hydroxyl or carbonyls) in the incorrect position, ligands were aligned using SMARTS strings enclosing the key functional group. In the case of 1-chloroheptane, the reference ligand 1-octanol was copied and manually mutated by converting the hydroxyl oxygen to a chlorine atom and removing the terminal methyl group. The resulting structure was then exported to a `_pv.maegz` file which is compatible with FEP+. We performed 20ns of FEP simulations for each system using  $\mu$ VT ensemble and 20 lambda windows on 4 GPUs, each using scripts containing modified versions of the following command:

```
"${SCHRODINGER}/fep_plus" -HOST localhost -SUBHOST <gpu-host> -ppj 4 -ffbuilder  
-ff-host <gpu-host> -time 20000.0 -ensemble muVT -seed 2007 \  
-membrane -water TIP4PEW -lambda_windows 20 -JOBNAME <job-name> \  
<system_pv.maegz>
```

## Protein RBFEP

The relaxed membrane system produced in the above “Membrane Generation and Relaxation” for each membrane lipid was subsequently converted to a pose viewer object compatible with FEP+ using the command:

```
$SCHRODINGER/run membrane_cms2fep.py <relaxed-membrane-system.cms>
```

We then ran 25ns protein RBFEP simulations using scripts containing the following command:

```
$SCHRODINGER/fep_plus -HOST localhost -SUBHOST gpu-a4500 -ppj 4 \  
-time 25000 -membrane -water TIP4PEW -ensemble NVT \  
-protein mutations.txt -solvent_as1 "(NOT (chain.name A AND res.ptype  
EOL))" \  
-JOBNAME <job-name> <system_pv.mae>
```

where the mutations.txt file contains the mutations described above.

## Chapter 6: Conclusions and Future Directions

### 6.1 Significance of research

In this thesis, we have described the development and use of methods to study protein-compound interactions at several scales. In chapters 2 and 3, we described efforts to expand the range of compounds which can be considered by PrePCI, our structure-based, proteome-scale PCI prediction method, by integrating chemoinformatic chemical similarity scores with rapid protein structural and sequence-based comparisons. These expansions have increased our coverage of chemical space by over two orders of magnitude while benchmarking results indicate performance is comparable to other similar methods. Such predictions can form the basis of expanded interactomes for systems biology studies, evaluating the impact of altered metabolite states or drug perturbation on biological networks.

Secondly, our preliminary work evaluating the structural specificity of druggable protein-protein interaction pockets provides a worst case estimate for the prospects of designing selective PPI inhibitors. While druggable PPI pockets in tyrosine kinase, ras kinase and trypsin domains were associated with fewer predicted off-targets relative to comparable pockets in active sites, the same was not observed in SET domains, nuclear hormone receptor ligand binding domains, or kinases generally. These results suggest that the success of developing successfully highly specific PPI inhibitors may be family dependent. Nevertheless, the presence of sampling bias, holostructure bias and methodological bias described in chapter 4 leave room for further work to clarify these results. We describe possible approaches to mitigate these sources of bias in the following section.

Finally, our calculations of MhOR5 affinity with a range of retrospective compounds shows that FEP+ can reliably predict the relative affinities of compounds which bind to MhOR5 while our prospective predictions for new compounds provide testable hypotheses of compounds which MhOR5 can likely recognize. In contrast, our experiment with lipid membrane composition did not resolve the FEP prediction inaccuracies relative to the experimental affinity measurements. Further experimental and computational work is required to optimize the system and ensure reliable estimation of mutant MhOR5 affinities with eugenol.

## **6.2 Future Directions**

### **6.2.1 Future Directions for PrePCI**

Earlier in this thesis, we demonstrated that the current PrePCI pipeline performs comparably to existing methods while covering increasing portions of chemical space and retaining proteome-scale. However, there are still numerous ways that the PrePCI algorithm could be improved or modifications that could be made. Here we list several possible modifications which could be attempted to further improve PrePCI.

#### **6.2.1.1 Improvements of PrePCI pipeline**

Recent developments in deep learning have led to rapid structure alignment tools which can dramatically accelerate the search for structurally similar template proteins. While we have traditionally used *ska*<sup>97</sup> to perform pairwise structure alignments, replacing *ska* with newly developed machine-learning methods like *FoldSeek*<sup>220</sup> could enable more rapid searching of protein space. The use of *Foldseek* could potentially help identify additional similarity relationships between template and query proteins as well as facilitate the use of more query

structures in our model database, thereby allowing us to sample protein space more fully and identify protein conformations which may be more conducive to ligand binding.

Additionally, the sequence score was based entirely on the global sequence alignment of the template protein to the human proteome. Instead, the approach used by Homolobind<sup>221</sup>, in which the sequence similarity of residues aligned to contact residues in the PDB template could be used to score the similarity of the putative interface specifically, rather than the protein sequence as a whole.

Another potential modification could be to replace the LT-scanner scoring function. For example, a rapid pose-scoring method like RF-score<sup>63,64</sup>, could enable PrePCI to attempt to predict affinities rather than binary classifications. It is likely that any pose-scoring method used would need to be retrained specifically on LT-scanner interaction models to account for the crudeness of unoptimized structural alignments.

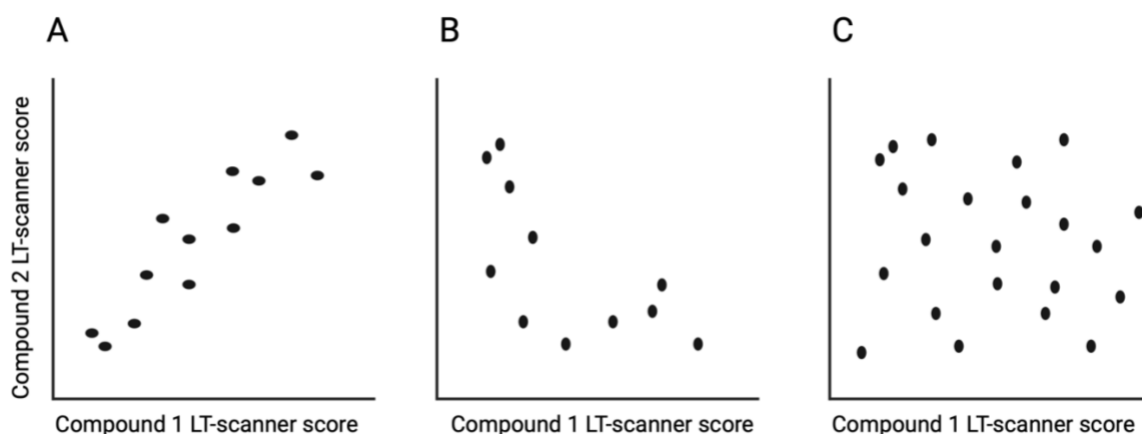
For the chemical similarity component, there are various ways to potentially improve the similarity calculations. One approach would be to simply replace the Tanimoto Coefficient with the Tversky Index, using the PDB compound as the reference compound. Such an approach could in principle improve the specificity of predictions by prioritizing the reference ligand's fragments rather than treating the two compounds identically. Alternatively, it may be possible to develop fingerprints based on the chemical fragments which form intermolecular interactions with the protein. Chemical similarity calculations, either the Tanimoto Coefficient or Tversky index, could then be computed specifically based on the fragments which contact the protein to assess the degree to which the query ligand is capable of recapitulating the contacts made by the reference ligand.

Finally, as I discussed in Chapter 3, the structural superpositions generated by LT-scanner are generally fairly crude and can lead to clashes between the ligand and the protein. Use of PrePCI predictions in systems biology applications would likely benefit by reducing the prevalence of false positives, thereby improving the specificity of the predictions and derived networks. Therefore, it may be useful to perform refinements of predicted PCI models for select PCIs of interest. For example, automated restrained minimization or Monte-Carlo sampling of the protein configuration in the presence of the ligand may enable a more realistic assessment of whether the ligand binding site can accommodate the small molecule. While this would likely be infeasible on the scale of PrePCI, it may be possible for the 400 compounds in the MIDAS metabolite reference set<sup>26</sup> or even in the more expansive Human Metabolome Database<sup>222</sup> which contains closer to 3000 metabolites. Such analysis could further improve our confidence PrePCI's predictions and potentially enable a more thorough integration of PrePCI predictions with systems biology applications.

#### **6.2.1.2 Application of PrePCI interactomes**

In addition to possible improvements, there are several possible applications of PrePCI in the analysis of, and hypothesis generation for, biological systems. One intriguing possibility is the application of PrePCI to detect allostery by examining the correlation of LT-scanner scores among pairs of ligands. Because LT-scanner provides a distinct score for each model it evaluates, it directly links the specific structural model to the binding prediction. Use of multiple different models would generate an array of scores for each ligand predicted to bind the protein domain which, if there are sufficiently large fluctuations between models, could include fairly large changes in the associated LT-scanner score. By simultaneously examining the LT-scanner scores of pairs of ligands which bind at distinct sites, it may be possible to identify pairs of ligands which

exhibit positive correlation, where binding of one implies binding to the other, or conversely, inverse correlation, where binding of one is negatively associated with the other. It is possible that significant correlations could be identified by inferring whether expected distributions, such as uniform, positive correlation or negative correlation (see Figure 6.1), are consistent with the observed datapoints using goodness-of-fit tests. Alternatively, it is possible that the distributions corresponding to positive, negative and non-allosteric modulators could be learned using machine learning.



**Figure 6.1 Possible Application of PrePCI to identify allosteric relationships between ligands for a given protein.**

LT-scanner provides a unique score associating each protein model to a small molecule. By examining the correlation LT-scanner scores for different small molecules with the same protein model it may be possible to identify compounds which bind (A) cooperatively (B) competitively or (C) independently of one another.

Additionally, I presented several hundred predictions of small molecules predicted to bind probable master regulator binding proteins within the PPI interface, specifically focusing on cases with low sequence similarity to a template. These predictions could easily be expanded systematically to include more of the protein-protein interactome and cases of biological or medicinal interest could be more rigorously evaluated using the methods described in chapter 3.

## 6.2.2 Alternative approaches to analyzing druggable pockets in PPI interfaces

In the discussion in Chapter 4, I described several potential sources of bias within the datasets used to evaluate the family-level similarity of druggable PPI pockets compared to typical ligandable binding sites. Below are several recommendations for strategies to mitigate these biases and interrogate the selectivity of druggable PPI pockets more extensively.

### Sampling bias

The uneven representation of different proteins within the PDB likely results in insufficient sampling of the structural space available to each pocket in some proteins relative to their more popular homologs. Proteins which are underrepresented in the PDB could be supplemented using homology models to more comprehensively sample the structural space of each PPI pocket. Comprehensive generation of models where each PDB file in the scPDB/PPIome dataset is used as a template for every other protein in the family could enable a more complete sampling of the conformational space for each protein and in turn provide a more robust sense of how often pockets are shared among proteins within a family.

### Holostructure bias

Furthermore, I noted that comparisons between PPIome pockets and scPDB pockets were complicated by the presence of ligands in the scPDB holo-structures. The presence of ligands in a protein structure can lead to the formation of specific intermolecular bonds which constrain the geometry of the protein structure relative to the apo-structure. Accordingly, it is possible that the pockets within scPDB structures were, at least in part, more similar to one another because they were predominantly present in the holo-state, rather than the apo-state as PPIome proteins were. It would be interesting to repeat the analysis, replacing scPDB structures with apo-structures of the

same proteins, to see if the absence of ligands results in pockets with greater inter-protein variability and places the scPDB protein pocket similarity, and the associated number of off-target proteins, more on par with the PPIome.

### Methodological bias

Finally, in the analysis described above, pockets for individual proteins were clustered based on their structural similarity. Off-target proteins for a given protein cluster were identified on the basis of the protein's most similar pocket relative to the cluster. This approach was chosen on the expectation that, if a pocket within the cluster were targeted by a drug, then the second protein's most similar pocket would be the most likely off-target of the drug. Such an approach suggests an upper bound on the plausibility of identifying selective inhibitors, as it identifies whether other proteins within the family share a similar pocket. However, it may not be optimal for determining whether the potential for selective inhibitors exists, as each protein could sample additional low-energy conformations yielding structurally distinct pockets which could be pharmacologically targeted. Accordingly, an alternative approach to address this concern of identifying unique pockets could be as follows. 1) Rotate all template PDB files to a single coordinate frame using ska. 2) Generate homology models for each member of a family using each of the PDB structures as templates for each protein in the family, thereby equally sampling structure space for all proteins. 3) Identify pockets using Volsite (or other methods cited in chapter 4). 4) Group the pockets for each individual protein by spatial clustering of pocket geometric centers. 5) Create correspondences between the pockets of each protein based on spatial proximity. This ensures that subsequent pocket comparisons are only made between analogous pockets and that low pocket similarity indicates large local structural fluctuations, rather than simply comparing two unrelated pockets. 6) For pockets associated with each unique site, identify whether

highly dissimilar pockets between proteins exist, either by evaluating the Tversky index of the least similar pocket, or more modestly the similarity of the 25<sup>th</sup> percentile as suggested by Guo and Chen<sup>166</sup> to mitigate the impact of spurious outlier pockets. Such an approach may more efficiently identify structurally distinct pockets which could be uniquely targeted to disrupt specific protein PPI interfaces.

### **6.2.3 Additional analyses of MhOR5**

Additional MhOR5 dose-response activity measurements for the prospective ligands studied in chapter 5 will be performed by the Ruta group to assess the accuracy of our prospective FEP predictions.

The choice of membrane lipid composition was unable to resolve the discrepancies between experimentally observed affinity measurements and binding free energies predicted by FEP+. Moving forward, we could try generating more realistic membranes, for example by including cholesterol molecules or by adding mixed and asymmetric membrane lipid distributions. Such approaches are possible using CHARMM-GUI<sup>223</sup>, but the resulting structures are incompatible with Desmond, FEP+ and the Schrodinger Suite of tools and would prompt the use of alternative free energy calculation engines. Alternatively, we had previously noticed that the helices constituting the walls of the binding pocket appeared more stable using AMBER forcefield parameters. While it is not currently possible to use AMBER forcefield parameters within FEP+, we are discussing how these parameters might be imported into FEP+ and enable free energy calculations with additional forcefields which may be better suited to the MhOR5 system. Finally, it is possible that the mutation of residues M209 and I213 dramatically affect the MhOR5's structure in a manner which is not captured in the simulations. Experimental determination of these

mutant structures could help explain the persistent discrepancies between the simulation and the experimental results.

## Appendix: Code Locations and Information

This table contains the location of code/resources associated with PrePCI and scPDB/PPIome pocket analysis located on the C2B2 cluster and a brief description of contents.

Resource	Location	Description
PrePCI Code	/ifs/home/c2b2/bh_lab/shares/cvsroot/hfpd /prepci	Central location of PrePCI code. Accessible using the Concurrent Versions System (CVS) with command “cvs checkout hfpd” rather than accessing directly.
PrePCI Utilities	/ifs/data/c2b2/bh_lab/shares/prepci/utilities	General directory containing utilities associated with Protein Science 2023 Paper
Chemical Similarity Files	/ifs/data/c2b2/bh_lab/shares/prepci/utilities /chemical_comparison/pubchem	Directory containing chemical similarity comparison files, each file corresponds to an individual PDB compound and lists all similar Pubchem compounds and associated Tanimoto Coefficients
Bioactivity Data	/ifs/data/c2b2/bh_lab/shares/prepci/utilities /experimental_results/protein_science_2023/pubchem_pcis	Directory containing high-throughput experimental bioactivity data from Pubchem Described in Chapter 2. Each file corresponds to an individual protein and contains all identified PCIs
Holo-structure list	/ifs/data/c2b2/bh_lab/shares/prepci/utilities /holostructure_templates	List of holo-structures used as templates in Protein Science 2023 PrePCI paper
Skads Databases	/ifs/data/c2b2/bh_lab/shares/prepci/utilities /holostructure_templates/skads	Directory containing the skads databases used in 2023 Protein Science PrePCI paper
Benchmarking Results	/ifs/data/c2b2/bh_lab/shares/prepci/utilities /protein_science_2023_benchmarking/benchmarking.tar.gz	Zip file containing benchmarking results on Yamanishi, DUDE and DEKOIS datasets.
Website Databases	/ifs/data/c2b2/bh_lab/shares/prepci/utilities /website/Databases/databases	Contains flat files underlying mysql table used for PrePCI website
Website HTML Directory	/ifs/home/www/vhosts/honiglab.c2b2.colu mbia.edu/html	Directory containing main HTML for PrePCI website (prepci.html, prepci-help.html)
PrePCI LT-scanner/Sequence PCI Result Page	/ifs/home/www/vhosts/honiglab.c2b2.colu mbia.edu/html/WEB_DB/cgi- bin/prepci_v2_pci_result.cgi	CGI perl script which recalls and displays LT-scanner/Sequence results
PrePCI Similar Compounds Result Page	/ifs/home/www/vhosts/honiglab.c2b2.colu mbia.edu/html/WEB_DB/cgi- bin/prepci_v2_simcompd_result.cgi	CGI perl script which identifies compounds similar to a PDB compound of interest and generates html to display them
PCIs predicted to bind at master regulator PPI interface	/ifs/scratch/c2b2/bh_lab/shares/steve/ppi_ pci_overlapping_interfaces/ltscanner_prep pi_interfaces/scripts/moma	Directory containing the scripts and results of PCIs predicted to bind at master regulators PPI interfaces

Pocket Analysis Directory	/ifs/scratch/c2b2/bh_lab/shares/steve/ppi_pci_overlapping_interfaces/ichem/redo_from_scratch	Directory containing the code, utilities and results associated with scPDB and PPIome pocket analysis
Software directory	/ifs/scratch/c2b2/bh_lab/shares/steve/software	Directory containing miscellaneous third-party software used in scPDB and PPIome pocket analysis (ie chimera, openeye, shaper, ichem (Volsite))

This table contains the location of code/resources associated with the MhOR5 project located on the gates cluster and a brief description of contents. Each directory contains README and NOTES files describing the directory's contents and procedures used to generate results.

Resource	Location	Description
Prospective Ligand FEP	/mnt/beegfs/home/honig/sjt2151/MhOR5/prospective_fep	Directory containing scripts and results used in prospective ligand FEP. Results used in thesis primarily in "membrane_preparation_pierre_ifd-md" subdirectory. Additionally, "scripts" subdirectory contains commonly used scripts
Prospective Ligand Membrane Generation and Relaxation	/mnt/beegfs/home/honig/sjt2151/MhOR5/prospective_fep/membrane_preparation_pierre_ifd-md	Contains intermediate files associated with generating the optimized POPC membrane for each of the initial IFD-MD starting poses (subdirectories labeled by reference ligand)
Prospective Ligand Membrane Generation and Relaxation	/mnt/beegfs/home/honig/sjt2151/MhOR5/prospective_fep/membrane_preparation_pierre_ifd-md/fep	Contains intermediate files associated with running prospective and retrospective RBFEP for each of the congeneric series (subdirectories labeled by series name and reference ligand)
Protein Mutation FEP	/mnt/beegfs/home/honig/sjt2151/MhOR5/protein_fep	Directory containing scripts and results used in protein FEP analysis. Contains two main subdirectories, "scripts" containing commonly used scripts, and "different_membranes" containing the results of protein FEP using all four membrane lipids compatible with FEP+

## References

1. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet.* 2004;5(2):101-13. doi: 10.1038/nrg1272. PubMed PMID: 14735121.
2. Walhout AJ, Boulton SJ, Vidal M. Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast.* 2000;17(2):88-94. doi: 10.1002/1097-0061(20000630)17:2<88::AID-YEA20>3.0.CO;2-Y. PubMed PMID: 10900455; PMCID: PMC2448329.
3. Bruckner A, Polge C, Lentze N, Auerbach D, Schlattner U. Yeast two-hybrid, a powerful tool for systems biology. *Int J Mol Sci.* 2009;10(6):2763-88. Epub 20090618. doi: 10.3390/ijms10062763. PubMed PMID: 19582228; PMCID: PMC2705515.
4. Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, Bian W, Brignall R, Cafarelli T, Campos-Laborie FJ, Charlotiaux B, Choi D, Cote AG, Daley M, Deimling S, Desbuleux A, Dricot A, Gebbia M, Hardy MF, Kishore N, Knapp JJ, Kovacs IA, Lemmens I, Mee MW, Mellor JC, Pollis C, Pons C, Richardson AD, Schlabach S, Teeking B, Yadav A, Babor M, Balcha D, Basha O, Bowman-Colin C, Chin SF, Choi SG, Colabella C, Coppin G, D'Amata C, De Ridder D, De Rouck S, Duran-Frigola M, Ennajdaoui H, Goebels F, Goehring L, Gopal A, Haddad G, Hatchi E, Helmy M, Jacob Y, Kassa Y, Landini S, Li R, van Lieshout N, MacWilliams A, Markey D, Paulson JN, Rangarajan S, Rasla J, Rayhan A, Rolland T, San-Miguel A, Shen Y, Sheykhkarimli D, Sheynkman GM, Simonovsky E, Tasan M, Tejada A, Tropepe V, Twizere JC, Wang Y, Weatheritt RJ, Weile J, Xia Y, Yang X, Yeger-Lotem E, Zhong Q, Aloy P, Bader GD, De Las Rivas J, Gaudet S, Hao T, Rak J, Tavernier J, Hill DE, Vidal M, Roth FP, Calderwood MA. A reference map of the human binary protein interactome. *Nature.* 2020;580(7803):402-8. Epub 20200408. doi: 10.1038/s41586-020-2188-x. PubMed PMID: 32296183; PMCID: PMC7169983.
5. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature.* 2002;415(6868):180-3. doi: 10.1038/415180a. PubMed PMID: 11805837.
6. Alonso-Lopez D, Campos-Laborie FJ, Gutierrez MA, Lambourne L, Calderwood MA, Vidal M, De Las Rivas J. APID database: redefining protein-protein interaction experimental evidences and binary interactomes. *Database (Oxford).* 2019;2019. Epub 20190101. doi: 10.1093/database/baz005. PubMed PMID: 30715274; PMCID: PMC6354026.
7. Bulyk ML. Analysis of sequence specificities of DNA-binding proteins with protein binding microarrays. *Methods Enzymol.* 2006;410:279-99. doi: 10.1016/S0076-6879(06)10013-0. PubMed PMID: 16938556; PMCID: PMC2747587.
8. Mardis ER. ChIP-seq: welcome to the new frontier. *Nat Methods.* 2007;4(8):613-4. doi: 10.1038/nmeth0807-613. PubMed PMID: 17664943.

9. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol.* 2008;26(12):1351-9. Epub 20081116. doi: 10.1038/nbt.1508. PubMed PMID: 19029915; PMCID: PMC2597701.
10. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* 2008;5(7):621-8. Epub 20080530. doi: 10.1038/nmeth.1226. PubMed PMID: 18516045.
11. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009;10(1):57-63. doi: 10.1038/nrg2484. PubMed PMID: 19015660; PMCID: PMC2949280.
12. Haas BJ, Zody MC. Advancing RNA-Seq analysis. *Nat Biotechnol.* 2010;28(5):421-3. doi: 10.1038/nbt0510-421. PubMed PMID: 20458303.
13. Dephoure N, Gould KL, Gygi SP, Kellogg DR. Mapping and analysis of phosphorylation sites: a quick guide for cell biologists. *Mol Biol Cell.* 2013;24(5):535-42. doi: 10.1091/mbc.E12-09-0677. PubMed PMID: 23447708; PMCID: PMC3583658.
14. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. *Nat Genet.* 2005;37(4):382-90. Epub 20050320. doi: 10.1038/ng1532. PubMed PMID: 15778709.
15. Wang K, Saito M, Bisikirska BC, Alvarez MJ, Lim WK, Rajbhandari P, Shen Q, Nemenman I, Basso K, Margolin AA, Klein U, Dalla-Favera R, Califano A. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nat Biotechnol.* 2009;27(9):829-39. Epub 20090909. doi: 10.1038/nbt.1563. PubMed PMID: 19741643; PMCID: PMC2753889.
16. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, Honig B. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature.* 2012;490(7421):556-60. Epub 20120930. doi: 10.1038/nature11503. PubMed PMID: 23023127; PMCID: PMC3482288.
17. Giorgi FM, Lopez G, Woo JH, Bisikirska B, Califano A, Bansal M. Inferring protein modulation from gene expression data using conditional mutual information. *PLoS One.* 2014;9(10):e109569. Epub 20141014. doi: 10.1371/journal.pone.0109569. PubMed PMID: 25314274; PMCID: PMC4196905.
18. Garzon JI, Deng L, Murray D, Shapira S, Petrey D, Honig B. A computational interactome and functional annotation for the human proteome. *Elife.* 2016;5. Epub 20161022. doi: 10.7554/eLife.18715. PubMed PMID: 27770567; PMCID: PMC5115866.
19. Lasso G, Mayer SV, Winkelmann ER, Chu T, Elliot O, Patino-Galindo JA, Park K, Rabadan R, Honig B, Shapira SD. A Structure-Informed Atlas of Human-Virus Interactions. *Cell.* 2019;178(6):1526-41 e16. Epub 20190829. doi: 10.1016/j.cell.2019.08.005. PubMed PMID: 31474372; PMCID: PMC6736651.
20. Silverbush D, Sharan R. A systematic approach to orient the human protein-protein interaction network. *Nat Commun.* 2019;10(1):3015. Epub 20190709. doi: 10.1038/s41467-019-10887-6. PubMed PMID: 31289271; PMCID: PMC6617457.
21. Kovacs IA, Luck K, Spirohn K, Wang Y, Pollis C, Schlabach S, Bian W, Kim DK, Kishore N, Hao T, Calderwood MA, Vidal M, Barabasi AL. Network-based prediction of protein interactions. *Nat Commun.* 2019;10(1):1240. Epub 20190318. doi: 10.1038/s41467-019-09177-y. PubMed PMID: 30886144; PMCID: PMC6423278.
22. Broyde J, Simpson DR, Murray D, Paull EO, Chu BW, Tagore S, Jones SJ, Griffin AT, Giorgi FM, Lachmann A, Jackson P, Sweet-Cordero EA, Honig B, Califano A. Oncoprotein-

- specific molecular interaction maps (SigMaps) for cancer network analyses. *Nat Biotechnol.* 2021;39(2):215-24. Epub 20200914. doi: 10.1038/s41587-020-0652-7. PubMed PMID: 32929263; PMCID: PMC7878435.
23. Petrey D, Zhao H, Trudeau SJ, Murray D, Honig B. PrePPI: A Structure Informed Proteome-wide Database of Protein-Protein Interactions. *J Mol Biol.* 2023;435(14):168052. Epub 20230317. doi: 10.1016/j.jmb.2023.168052. PubMed PMID: 36933822; PMCID: PMC10293085.
24. Rosenberger G, Li W, Turunen M, He J, Subramaniam PS, Pampou S, Griffin AT, Karan C, Kerwin P, Murray D, Honig B, Liu Y, Califano A. Network-based elucidation of colon cancer drug resistance by phosphoproteomic time-series analysis. *bioRxiv.* 2023. Epub 20230216. doi: 10.1101/2023.02.15.528736. PubMed PMID: 36824919; PMCID: PMC9949144.
25. Piazza I, Kochanowski K, Cappelletti V, Fuhrer T, Noor E, Sauer U, Picotti P. A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication. *Cell.* 2018;172(1-2):358-72 e23. Epub 20180104. doi: 10.1016/j.cell.2017.12.006. PubMed PMID: 29307493.
26. Hicks KG, Cluntun AA, Schubert HL, Hackett SR, Berg JA, Leonard PG, Ajalla Aleixo MA, Zhou Y, Bott AJ, Salvatore SR, Chang F, Blevins A, Barta P, Tilley S, Leifer A, Guzman A, Arok A, Fogarty S, Winter JM, Ahn HC, Allen KN, Block S, Cardoso IA, Ding J, Dreveny I, Gasper WC, Ho Q, Matsuura A, Palladino MJ, Prajapati S, Sun P, Tittmann K, Tolan DR, Unterlass J, VanDemark AP, Vander Heiden MG, Webb BA, Yun CH, Zhao P, Wang B, Schopfer FJ, Hill CP, Nonato MC, Muller FL, Cox JE, Rutter J. Protein-metabolite interactomics of carbohydrate metabolism reveal regulation of lactate dehydrogenase. *Science.* 2023;379(6636):996-1003. Epub 20230309. doi: 10.1126/science.abm3452. PubMed PMID: 36893255; PMCID: PMC10262665.
27. Kemp RG, Foe LG. Allosteric regulatory properties of muscle phosphofructokinase. *Mol Cell Biochem.* 1983;57(2):147-54. doi: 10.1007/BF00849191. PubMed PMID: 6228716.
28. Vilar JM, Guet CC, Leibler S. Modeling network dynamics: the lac operon, a case study. *J Cell Biol.* 2003;161(3):471-6. doi: 10.1083/jcb.200301125. PubMed PMID: 12743100; PMCID: PMC2172934.
29. Overduin M, Kervin TA. The phosphoinositide code is read by a plethora of protein domains. *Expert Rev Proteomics.* 2021;18(7):483-502. Epub 20210823. doi: 10.1080/14789450.2021.1962302. PubMed PMID: 34351250.
30. Pemberton JG, Balla T. Polyphosphoinositide-Binding Domains: Insights from Peripheral Membrane and Lipid-Transfer Proteins. *Adv Exp Med Biol.* 2019;1111:77-137. doi: 10.1007/5584\_2018\_288. PubMed PMID: 30483964; PMCID: PMC8284841.
31. Matthews BW. X-ray Crystallographic Studies of Proteins. *Ann Rev Phys Chem.* 1976;27:493-523.
32. Zhang X, Jin L, Fang Q, Hui WH, Zhou ZH. 3.3 A cryo-EM structure of a nonenveloped virus reveals a priming mechanism for cell entry. *Cell.* 2010;141(3):472-82. Epub 20100415. doi: 10.1016/j.cell.2010.03.041. PubMed PMID: 20398923; PMCID: PMC3422562.
33. Liu H, Jin L, Koh SB, Atanasov I, Schein S, Wu L, Zhou ZH. Atomic structure of human adenovirus by cryo-EM reveals interactions among protein networks. *Science.* 2010;329(5995):1038-43. doi: 10.1126/science.1187433. PubMed PMID: 20798312; PMCID: PMC3412078.
34. Palmer AG, 3rd. NMR characterization of the dynamics of biomacromolecules. *Chem Rev.* 2004;104(8):3623-40. doi: 10.1021/cr030413t. PubMed PMID: 15303831.

35. Freyer MW, Lewis EA. Isothermal titration calorimetry: experimental design, data analysis, and probing macromolecule/ligand binding and kinetic interactions. *Methods Cell Biol.* 2008;84:79-113. doi: 10.1016/S0091-679X(07)84004-0. PubMed PMID: 17964929.
36. Stahelin RV. Surface plasmon resonance: a useful technique for cell biologists to characterize biomolecular interactions. *Mol Biol Cell.* 2013;24(7):883-6. doi: 10.1091/mbc.E12-10-0713. PubMed PMID: 23533209; PMCID: PMC3608497.
37. Orsak T, Smith TL, Eckert D, Lindsley JE, Borges CR, Rutter J. Revealing the allosterome: systematic identification of metabolite-protein interactions. *Biochemistry.* 2012;51(1):225-32. Epub 20111209. doi: 10.1021/bi201313s. PubMed PMID: 22122470.
38. Lomenick B, Hao R, Jonai N, Chin RM, Aghajan M, Warburton S, Wang J, Wu RP, Gomez F, Loo JA, Wohlschlegel JA, Vondriska TM, Pelletier J, Herschman HR, Clardy J, Clarke CF, Huang J. Target identification using drug affinity responsive target stability (DARTS). *Proc Natl Acad Sci U S A.* 2009;106(51):21984-9. Epub 20091207. doi: 10.1073/pnas.0910040106. PubMed PMID: 19995983; PMCID: PMC2789755.
39. Cappelletti V, Hauser T, Piazza I, Pepelnjak M, Malinowska L, Fuhrer T, Li Y, Dorig C, Boersema P, Gillet L, Grossbach J, Dugourd A, Saez-Rodriguez J, Beyer A, Zamboni N, Caflisch A, de Souza N, Picotti P. Dynamic 3D proteomes reveal protein functional alterations at high resolution in situ. *Cell.* 2021;184(2):545-59 e22. Epub 20201223. doi: 10.1016/j.cell.2020.12.021. PubMed PMID: 33357446; PMCID: PMC7836100.
40. Feng Y, De Franceschi G, Kahraman A, Soste M, Melnik A, Boersema PJ, de Laureto PP, Nikolaev Y, Oliveira AP, Picotti P. Global analysis of protein structural changes in complex proteomes. *Nat Biotechnol.* 2014;32(10):1036-44. Epub 20140914. doi: 10.1038/nbt.2999. PubMed PMID: 25218519.
41. Piazza I, Beaton N, Bruderer R, Knobloch T, Barbisan C, Chandat L, Sudau A, Siepe I, Rinner O, de Souza N, Picotti P, Reiter L. A machine learning-based chemoproteomic approach to identify drug targets and binding sites in complex proteomes. *Nat Commun.* 2020;11(1):4200. Epub 20200821. doi: 10.1038/s41467-020-18071-x. PubMed PMID: 32826910; PMCID: PMC7442650.
42. West GM, Tang L, Fitzgerald MC. Thermodynamic analysis of protein stability and ligand binding using a chemical modification- and mass spectrometry-based strategy. *Anal Chem.* 2008;80(11):4175-85. Epub 20080506. doi: 10.1021/ac702610a. PubMed PMID: 18457414.
43. Dearmond PD, Xu Y, Strickland EC, Daniels KG, Fitzgerald MC. Thermodynamic analysis of protein-ligand interactions in complex biological mixtures using a shotgun proteomics approach. *J Proteome Res.* 2011;10(11):4948-58. Epub 20110928. doi: 10.1021/pr200403c. PubMed PMID: 21905665; PMCID: PMC3208786.
44. Strickland EC, Geer MA, Tran DT, Adhikari J, West GM, DeArmond PD, Xu Y, Fitzgerald MC. Thermodynamic analysis of protein-ligand binding interactions in complex biological mixtures using the stability of proteins from rates of oxidation. *Nat Protoc.* 2013;8(1):148-61. Epub 20121220. doi: 10.1038/nprot.2012.146. PubMed PMID: 23257983; PMCID: PMC3717606.
45. Willett P. Similarity searching using 2D structural fingerprints. *Methods Mol Biol.* 2011;672:133-58. doi: 10.1007/978-1-60761-839-3\_5. PubMed PMID: 20838967.
46. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol Inform.* 2010;29(6-7):476-88. Epub 20100706. doi: 10.1002/minf.201000061. PubMed PMID: 27463326.

47. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol.* 2007;25(2):197-206. doi: 10.1038/nbt1284. PubMed PMID: 17287757.
48. Luco JM, Ferretti FH. QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives. *J Chem Inf Comput Sci.* 1997;37(2):392-401. doi: 10.1021/ci960487o. PubMed PMID: 9090857.
49. Jayaraj PB, Jain S. Ligand based virtual screening using SVM on GPU. *Comput Biol Chem.* 2019;83:107143. Epub 20191110. doi: 10.1016/j.compbiolchem.2019.107143. PubMed PMID: 31743833.
50. Burden FR, Ford MG, Whitley DC, Winkler DA. Use of automatic relevance determination in QSAR studies using Bayesian neural networks. *J Chem Inf Comput Sci.* 2000;40(6):1423-30. doi: 10.1021/ci000450a. PubMed PMID: 11128101.
51. Burden FR, Winkler DA. Robust QSAR models using Bayesian regularized neural networks. *J Med Chem.* 1999;42(16):3183-7. doi: 10.1021/jm980697n. PubMed PMID: 10447964.
52. Jayaraj PB, Ajay MK, Nufail M, Gopakumar G, Jaleel UC. GPURFSCREEN: a GPU based virtual screening tool using random forest classifier. *J Cheminform.* 2016;8:12. Epub 20160301. doi: 10.1186/s13321-016-0124-8. PubMed PMID: 26933453; PMCID: PMC4772510.
53. Jayaraj PB, Sanjay S, Raja K, Gopakumar G, Jaleel UC. Ligand Based Virtual Screening Using Self-organizing Maps. *Protein J.* 2022;41(1):44-54. Epub 20220113. doi: 10.1007/s10930-021-10030-9. PubMed PMID: 35022993.
54. Hu S, Chen P, Gu P, Wang B. A Deep Learning-Based Chemical System for QSAR Prediction. *IEEE J Biomed Health Inform.* 2020;24(10):3020-8. Epub 20200228. doi: 10.1109/JBHI.2020.2977009. PubMed PMID: 32142459.
55. Sakai M, Nagayasu K, Shibui N, Andoh C, Takayama K, Shirakawa H, Kaneko S. Prediction of pharmacological activities from chemical structures with graph convolutional neural networks. *Sci Rep.* 2021;11(1):525. Epub 20210112. doi: 10.1038/s41598-020-80113-7. PubMed PMID: 33436854; PMCID: PMC7803991.
56. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem.* 2004;47(7):1739-49. doi: 10.1021/jm0306430. PubMed PMID: 15027865.
57. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem.* 2006;49(21):6177-96. doi: 10.1021/jm051256o. PubMed PMID: 17034125.
58. Miller EB, Murphy RB, Sindhikara D, Borrelli KW, Grisewood MJ, Ranalli F, Dixon SL, Jerome S, Boyles NA, Day T, Ghanakota P, Mondal S, Rafi SB, Troast DM, Abel R, Friesner RA. Reliable and Accurate Solution to the Induced Fit Docking Problem for Protein-Ligand Binding. *J Chem Theory Comput.* 2021;17(4):2630-9. Epub 20210329. doi: 10.1021/acs.jctc.1c00136. PubMed PMID: 33779166.
59. Bottegoni G, Kufareva I, Totrov M, Abagyan R. Four-dimensional docking: a fast and accurate account of discrete receptor flexibility in ligand docking. *J Med Chem.* 2009;52(2):397-406. doi: 10.1021/jm8009958. PubMed PMID: 19090659; PMCID: PMC2662720.

60. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. 2010;31(2):455-61. doi: 10.1002/jcc.21334. PubMed PMID: 19499576; PMCID: PMC3041641.
61. Baskaran SG, Sharp TP, Sharp KA. Computational Graphics Software for Interactive Docking and Visualization of Ligand-Protein Complementarity. *J Chem Inf Model*. 2021;61(3):1427-43. Epub 20210303. doi: 10.1021/acs.jcim.0c01485. PubMed PMID: 33656873.
62. Guterres H, Park SJ, Jiang W, Im W. Ligand-Binding-Site Refinement to Generate Reliable Holo Protein Structure Conformations from Apo Structures. *J Chem Inf Model*. 2021;61(1):535-46. Epub 20201218. doi: 10.1021/acs.jcim.0c01354. PubMed PMID: 33337877; PMCID: PMC7856192.
63. Ballester PJ, Mitchell JB. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*. 2010;26(9):1169-75. Epub 20100317. doi: 10.1093/bioinformatics/btq112. PubMed PMID: 20236947; PMCID: PMC3524828.
64. Wojcikowski M, Ballester PJ, Siedlecki P. Performance of machine-learning scoring functions in structure-based virtual screening. *Sci Rep*. 2017;7:46710. Epub 20170425. doi: 10.1038/srep46710. PubMed PMID: 28440302; PMCID: PMC5404222.
65. Rayka M, Firouzi R. GB-score: Minimally designed machine learning scoring function based on distance-weighted interatomic contact features. *Mol Inform*. 2023;42(3):e2200135. Epub 20230201. doi: 10.1002/minf.202200135. PubMed PMID: 36722733.
66. Brown BP, Mendenhall J, Geanes AR, Meiler J. General Purpose Structure-Based Drug Discovery Neural Network Score Functions with Human-Interpretable Pharmacophore Maps. *J Chem Inf Model*. 2021;61(2):603-20. Epub 20210126. doi: 10.1021/acs.jcim.0c01001. PubMed PMID: 33496578; PMCID: PMC7903419.
67. Durrant JD, McCammon JA. NNScore 2.0: a neural-network receptor-ligand scoring function. *J Chem Inf Model*. 2011;51(11):2897-903. Epub 20111103. doi: 10.1021/ci2003889. PubMed PMID: 22017367; PMCID: PMC3225089.
68. Durrant JD, McCammon JA. NNScore: a neural-network-based scoring function for the characterization of protein-ligand complexes. *J Chem Inf Model*. 2010;50(10):1865-71. doi: 10.1021/ci100244v. PubMed PMID: 20845954; PMCID: PMC2964041.
69. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein-Ligand Scoring with Convolutional Neural Networks. *J Chem Inf Model*. 2017;57(4):942-57. Epub 20170411. doi: 10.1021/acs.jcim.6b00740. PubMed PMID: 28368587; PMCID: PMC5479431.
70. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. *Bioinformatics*. 2018;34(21):3666-74. doi: 10.1093/bioinformatics/bty374. PubMed PMID: 29757353; PMCID: PMC6198856.
71. Karimi M, Wu D, Wang Z, Shen Y. DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*. 2019;35(18):3329-38. doi: 10.1093/bioinformatics/btz111. PubMed PMID: 30768156; PMCID: PMC6748780.
72. Karimi M, Wu D, Wang Z, Shen Y. Explainable Deep Relational Networks for Predicting Compound-Protein Affinities and Contacts. *J Chem Inf Model*. 2021;61(1):46-66. Epub 20201221. doi: 10.1021/acs.jcim.0c00866. PubMed PMID: 33347301; PMCID: PMC7987499.

73. Jiang M, Li Z, Zhang S, Wang S, Wang X, Yuan Q, Wei Z. Drug-target affinity prediction using graph neural network and contact maps. *RSC Adv.* 2020;10(35):20701-12. Epub 20200601. doi: 10.1039/d0ra02297g. PubMed PMID: 35517730; PMCID: PMC9054320.
74. Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. GraphDTA: predicting drug-target binding affinity with graph neural networks. *Bioinformatics.* 2021;37(8):1140-7. doi: 10.1093/bioinformatics/btaa921. PubMed PMID: 33119053.
75. Perez-Nueno VI, Rabal O, Borrell JI, Teixido J. APIF: a new interaction fingerprint based on atom pairs and its application to virtual screening. *J Chem Inf Model.* 2009;49(5):1245-60. doi: 10.1021/ci900043r. PubMed PMID: 19364101.
76. Da C, Kireev D. Structural protein-ligand interaction fingerprints (SPLIF) for structure-based virtual screening: method and benchmark study. *J Chem Inf Model.* 2014;54(9):2555-61. Epub 20140820. doi: 10.1021/ci500319f. PubMed PMID: 25116840; PMCID: PMC4170813.
77. Desaphy J, Raimbaud E, Ducrot P, Rognan D. Encoding protein-ligand interaction patterns in fingerprints and graphs. *J Chem Inf Model.* 2013;53(3):623-37. Epub 20130306. doi: 10.1021/ci300566n. PubMed PMID: 23432543.
78. Lyu J, Wang S, Balius TE, Singh I, Levit A, Moroz YS, O'Meara MJ, Che T, Alga E, Tolmacheva K, Tolmachev AA, Shoichet BK, Roth BL, Irwin JJ. Ultra-large library docking for discovering new chemotypes. *Nature.* 2019;566(7743):224-9. Epub 20190206. doi: 10.1038/s41586-019-0917-9. PubMed PMID: 30728502; PMCID: PMC6383769.
79. Sadybekov AA, Sadybekov AV, Liu Y, Iliopoulos-Tsoutsouvas C, Huang XP, Pickett J, Houser B, Patel N, Tran NK, Tong F, Zvonok N, Jain MK, Savych O, Radchenko DS, Nikas SP, Petasis NA, Moroz YS, Roth BL, Makriyannis A, Katritch V. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature.* 2022;601(7893):452-9. Epub 20211215. doi: 10.1038/s41586-021-04220-9. PubMed PMID: 34912117; PMCID: PMC9763054.
80. Lim H, Poleksic A, Yao Y, Tong H, He D, Zhuang L, Meng P, Xie L. Large-Scale Off-Target Identification Using Fast and Accurate Dual Regularized One-Class Collaborative Filtering and Its Application to Drug Repurposing. *PLoS Comput Biol.* 2016;12(10):e1005135. Epub 20161007. doi: 10.1371/journal.pcbi.1005135. PubMed PMID: 27716836; PMCID: PMC5055357.
81. Lim H, Gray P, Xie L, Poleksic A. Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. *Sci Rep.* 2016;6:38860. Epub 20161213. doi: 10.1038/srep38860. PubMed PMID: 27958331; PMCID: PMC5153628.
82. Liu Y, Wu M, Miao C, Zhao P, Li XL. Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction. *PLoS Comput Biol.* 2016;12(2):e1004760. Epub 20160212. doi: 10.1371/journal.pcbi.1004760. PubMed PMID: 26872142; PMCID: PMC4752318.
83. Liu B, Papadopoulos D, Malliaros FD, Tsoumakas G, Papadopoulos AN. Multiple similarity drug-target interaction prediction with random walks and matrix factorization. *Brief Bioinform.* 2022;23(5). doi: 10.1093/bib/bbac353. PubMed PMID: 36070659.
84. Lim H, He D, Qiu Y, Krawczuk P, Sun X, Xie L. Rational discovery of dual-indication multi-target PDE/Kinase inhibitor for precision anti-cancer therapy using structural systems pharmacology. *PLoS Comput Biol.* 2019;15(6):e1006619. Epub 20190617. doi: 10.1371/journal.pcbi.1006619. PubMed PMID: 31206508; PMCID: PMC6576746.
85. Li L, Koh CC, Reker D, Brown JB, Wang H, Lee NK, Liow HH, Dai H, Fan HM, Chen L, Wei DQ. Predicting protein-ligand interactions based on bow-pharmacological space and

- Bayesian additive regression trees. *Sci Rep.* 2019;9(1):7703. Epub 20190522. doi: 10.1038/s41598-019-43125-6. PubMed PMID: 31118426; PMCID: PMC6531441.
86. Meslamani J, Rognan D. Enhancing the accuracy of chemogenomic models with a three-dimensional binding site kernel. *J Chem Inf Model.* 2011;51(7):1593-603. Epub 20110621. doi: 10.1021/ci200166t. PubMed PMID: 21644501.
87. Shaikh N, Sharma M, Garg P. An improved approach for predicting drug-target interaction: proteochemometrics to molecular docking. *Mol Biosyst.* 2016;12(3):1006-14. doi: 10.1039/c5mb00650c. PubMed PMID: 26822863.
88. Ozturk H, Ozgur A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics.* 2018;34(17):i821-i9. doi: 10.1093/bioinformatics/bty593. PubMed PMID: 30423097; PMCID: PMC6129291.
89. Chen L, Tan X, Wang D, Zhong F, Liu X, Yang T, Luo X, Chen K, Jiang H, Zheng M. TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics.* 2020;36(16):4406-14. doi: 10.1093/bioinformatics/btaa524. PubMed PMID: 32428219.
90. Singh R, Sledzieski S, Bryson B, Cowen L, Berger B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proc Natl Acad Sci U S A.* 2023;120(24):e2220778120. Epub 20230608. doi: 10.1073/pnas.2220778120. PubMed PMID: 37289807; PMCID: PMC10268324.
91. Hwang H, Dey F, Petrey D, Honig B. Structure-based prediction of ligand-protein interactions on a genome-wide scale. *Proc Natl Acad Sci U S A.* 2017;114(52):13685-90. Epub 20171211. doi: 10.1073/pnas.1705381114. PubMed PMID: 29229851; PMCID: PMC5748165.
92. Zhou H, Cao H, Skolnick J. FINDSITE(comb2.0): A New Approach for Virtual Ligand Screening of Proteins and Virtual Target Screening of Biomolecules. *J Chem Inf Model.* 2018;58(11):2343-54. Epub 20181016. doi: 10.1021/acs.jcim.8b00309. PubMed PMID: 30278128; PMCID: PMC6437778.
93. Zhou H, Cao H, Skolnick J. FRAGSITE: A Fragment-Based Approach for Virtual Ligand Screening. *J Chem Inf Model.* 2021;61(4):2074-89. Epub 20210316. doi: 10.1021/acs.jcim.0c01160. PubMed PMID: 33724022; PMCID: PMC8243409.
94. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res.* 2000;28(1):235-42. doi: 10.1093/nar/28.1.235. PubMed PMID: 10592235; PMCID: PMC102472.
95. Da Silva F, Bret G, Teixeira L, Gonzalez CF, Rognan D. Exhaustive Repertoire of Druggable Cavities at Protein-Protein Interfaces of Known Three-Dimensional Structure. *J Med Chem.* 2019;62(21):9732-42. Epub 20191025. doi: 10.1021/acs.jmedchem.9b01184. PubMed PMID: 31603323.
96. Trudeau SJ, Hwang H, Mathur D, Begum K, Petrey D, Murray D, Honig B. PrePCI: A structure- and chemical similarity-informed database of predicted protein compound interactions. *Protein Sci.* 2023;32(4):e4594. doi: 10.1002/pro.4594. PubMed PMID: 36776141; PMCID: PMC10019447.
97. Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J Mol Biol.* 2000;301(3):665-78. doi: 10.1006/jmbi.2000.3973. PubMed PMID: 10966776.
98. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model.* 2010;50(5):742-54. doi: 10.1021/ci100050t. PubMed PMID: 20426451.

99. Bajusz D, Racz A, Heberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform.* 2015;7:20. Epub 20150520. doi: 10.1186/s13321-015-0069-3. PubMed PMID: 26052348; PMCID: PMC4456712.
100. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* 2019;47(D1):D1102-D9. doi: 10.1093/nar/gky1033. PubMed PMID: 30371825; PMCID: PMC6324075.
101. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* 2021;49(D1):D1388-D95. doi: 10.1093/nar/gkaa971. PubMed PMID: 33151290; PMCID: PMC7778930.
102. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596(7873):583-9. Epub 20210715. doi: 10.1038/s41586-021-03819-2. PubMed PMID: 34265844; PMCID: PMC8371605.
103. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Zidek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 2022;50(D1):D439-D44. doi: 10.1093/nar/gkab1061. PubMed PMID: 34791371; PMCID: PMC8728224.
104. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem.* 2012;55(14):6582-94. Epub 20120705. doi: 10.1021/jm300687e. PubMed PMID: 22716043; PMCID: PMC3405771.
105. Bauer MR, Ibrahim TM, Vogel SM, Boeckler FM. Evaluation and optimization of virtual screening workflows with DEKOIS 2.0--a public library of challenging docking benchmark sets. *J Chem Inf Model.* 2013;53(6):1447-62. Epub 20130612. doi: 10.1021/ci400115b. PubMed PMID: 23705874.
106. Adeshina YO, Deeds EJ, Karanicolas J. Machine learning classification can reduce false positives in structure-based virtual screening. *Proc Natl Acad Sci U S A.* 2020;117(31):18477-88. Epub 20200715. doi: 10.1073/pnas.2000585117. PubMed PMID: 32669436; PMCID: PMC7414157.
107. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics.* 2008;24(13):i232-40. doi: 10.1093/bioinformatics/btn162. PubMed PMID: 18586719; PMCID: PMC2718640.
108. Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, Martin M, Velankar S. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* 2019;47(D1):D482-D9. doi: 10.1093/nar/gky1114. PubMed PMID: 30445541; PMCID: PMC6324003.

109. Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, Oldfield TJ, O'Donovan C, Martin MJ, Kleywegt GJ. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.* 2013;41(Database issue):D483-9. Epub 20121129. doi: 10.1093/nar/gks1258. PubMed PMID: 23203869; PMCID: PMC3531078.
110. UniProt C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* 2023;51(D1):D523-D31. doi: 10.1093/nar/gkac1052. PubMed PMID: 36408920; PMCID: PMC9825514.
111. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* 2011;39(Database issue):D225-9. Epub 20101124. doi: 10.1093/nar/gkq1189. PubMed PMID: 21109532; PMCID: PMC3013737.
112. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2. PubMed PMID: 2231712.
113. Petrey D, Xiang Z, Tang CL, Xie L, Gimpelev M, Mitros T, Soto CS, Goldsmith-Fischman S, Kernytsky A, Schlessinger A, Koh IY, Alexov E, Honig B. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins.* 2003;53 Suppl 6:430-5. doi: 10.1002/prot.10550. PubMed PMID: 14579332.
114. Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods.* 2011;9(2):173-5. Epub 20111225. doi: 10.1038/nmeth.1818. PubMed PMID: 22198341.
115. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *J Cheminform.* 2011;3:33. Epub 20111007. doi: 10.1186/1758-2946-3-33. PubMed PMID: 21982300; PMCID: PMC3198950.
116. Weninger D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J Chem Inf Comput Sci.* 1988;28:31-6.
117. RDKit. RDKit.
118. Hu Y, Stumpfe D, Bajorath J. Advancing the activity cliff concept. *F1000Res.* 2013;2:199. Epub 20130930. doi: 10.12688/f1000research.2-199.v1. PubMed PMID: 24555097; PMCID: PMC3869489.
119. Paull EO, Aytes A, Jones SJ, Subramaniam PS, Giorgi FM, Douglass EF, Tagore S, Chu B, Vasciaveo A, Zheng S, Verhaak R, Abate-Shen C, Alvarez MJ, Califano A. A modular master regulator landscape controls cancer transcriptional identity. *Cell.* 2021;184(2):334-51 e20. Epub 20210111. doi: 10.1016/j.cell.2020.11.045. PubMed PMID: 33434495; PMCID: PMC8103356.
120. Xu T, Zhu K, Beutrait A, Vendome J, Borrelli KW, Abel R, Friesner RA, Miller EB. Induced-Fit Docking Enables Accurate Free Energy Perturbation Calculations in Homology Models. *J Chem Theory Comput.* 2022;18(9):5710-24. Epub 20220816. doi: 10.1021/acs.jctc.2c00371. PubMed PMID: 35972903.
121. Mobley DL, Gilson MK. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu Rev Biophys.* 2017;46:531-58. Epub 20170407. doi: 10.1146/annurev-biophys-070816-033654. PubMed PMID: 28399632; PMCID: PMC5544526.

122. Gallicchio E, Levy RM. Advances in all atom sampling methods for modeling protein-ligand binding affinities. *Curr Opin Struct Biol.* 2011;21(2):161-6. Epub 20110219. doi: 10.1016/j.sbi.2011.01.010. PubMed PMID: 21339062; PMCID: PMC3070828.
123. Wang J, Ishchenko A, Zhang W, Razavi A, Langley D. A highly accurate metadynamics-based Dissociation Free Energy method to calculate protein-protein and protein-ligand binding potencies. *Sci Rep.* 2022;12(1):2024. Epub 20220207. doi: 10.1038/s41598-022-05875-8. PubMed PMID: 35132139; PMCID: PMC8821539.
124. Zwanzig RW. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J Chem Phys.* 1954;22(8):1420-6.
125. Bennett CH. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics.* 1976;22:245-68.
126. Chen W, Cui D, Jerome SV, Michino M, Lenselink EB, Huggins DJ, Beautrait A, Vendome J, Abel R, Friesner RA, Wang L. Enhancing Hit Discovery in Virtual Screening through Absolute Protein-Ligand Binding Free-Energy Calculations. *J Chem Inf Model.* 2023;63(10):3171-85. Epub 20230511. doi: 10.1021/acs.jcim.3c00013. PubMed PMID: 37167486.
127. Sheng Z, Bimela JS, Wang M, Li Z, Guo Y, Ho DD. An optimized thermodynamics integration protocol for identifying beneficial mutations in antibody design. *Front Immunol.* 2023;14:1190416. Epub 20230519. doi: 10.3389/fimmu.2023.1190416. PubMed PMID: 37275896; PMCID: PMC10235760.
128. Shivakumar D, Harder E, Damm W, Friesner RA, Sherman W. Improving the Prediction of Absolute Solvation Free Energies Using the Next Generation OPLS Force Field. *J Chem Theory Comput.* 2012;8(8):2553-8. Epub 20120709. doi: 10.1021/ct300203w. PubMed PMID: 26592101.
129. Yang W, Pan Y, Zheng F, Cho H, Tai HH, Zhan CG. Free-energy perturbation simulation on transition states and redesign of butyrylcholinesterase. *Biophys J.* 2009;96(5):1931-8. doi: 10.1016/j.bpj.2008.11.051. PubMed PMID: 19254552; PMCID: PMC2717303.
130. Abel R, Wang L, Harder ED, Berne BJ, Friesner RA. Advancing Drug Discovery through Enhanced Free Energy Calculations. *Acc Chem Res.* 2017;50(7):1625-32. Epub 20170705. doi: 10.1021/acs.accounts.7b00083. PubMed PMID: 28677954.
131. Ioannidis H, Drakopoulos A, Tzitzoglaki C, Homeyer N, Kolarov F, Gkeka P, Freudenberger K, Liolios C, Gauglitz G, Cournia Z, Gohlke H, Kolocouris A. Alchemical Free Energy Calculations and Isothermal Titration Calorimetry Measurements of Aminoadamantanes Bound to the Closed State of Influenza A/M2TM. *J Chem Inf Model.* 2016;56(5):862-76. Epub 20160509. doi: 10.1021/acs.jcim.6b00079. PubMed PMID: 27105206.
132. Sergeeva AP, Katsamba PS, Liao J, Sampson JM, Bahna F, Mannepalli S, Morano NC, Shapiro L, Friesner RA, Honig B. Free Energy Perturbation Calculations of Mutation Effects on SARS-CoV-2 RBD::ACE2 Binding Affinity. *J Mol Biol.* 2023;435(15):168187. Epub 20230622. doi: 10.1016/j.jmb.2023.168187. PubMed PMID: 37355034; PMCID: PMC10286572.
133. Clark AJ, Gindin T, Zhang B, Wang L, Abel R, Murrett CS, Xu F, Bao A, Lu NJ, Zhou T, Kwong PD, Shapiro L, Honig B, Friesner RA. Free Energy Perturbation Calculation of Relative Binding Free Energy between Broadly Neutralizing Antibodies and the gp120 Glycoprotein of HIV-1. *J Mol Biol.* 2017;429(7):930-47. Epub 20161128. doi: 10.1016/j.jmb.2016.11.021. PubMed PMID: 27908641; PMCID: PMC5383735.

134. Ni C, Zheng K, Gao Y, Chen Y, Shi K, Ni C, Jin G, Yu G. ACOT4 accumulation via AKT-mediated phosphorylation promotes pancreatic tumorigenesis. *Cancer Lett.* 2021;498:19-30. Epub 20201024. doi: 10.1016/j.canlet.2020.09.022. PubMed PMID: 33148467.
135. Douse CH, Bloor S, Liu Y, Shamin M, Tchasovnikarova IA, Timms RT, Lehner PJ, Modis Y. Neuropathic MORC2 mutations perturb GHKL ATPase dimerization dynamics and epigenetic silencing by multiple structural mechanisms. *Nat Commun.* 2018;9(1):651. Epub 20180213. doi: 10.1038/s41467-018-03045-x. PubMed PMID: 29440755; PMCID: PMC5811534.
136. Wang G, Song Y, Liu T, Wang C, Zhang Q, Liu F, Cai X, Miao Z, Xu H, Xu H, Cao L, Li F. PAK1-mediated MORC2 phosphorylation promotes gastric tumorigenesis. *Oncotarget.* 2015;6(12):9877-86. doi: 10.18632/oncotarget.3185. PubMed PMID: 25888627; PMCID: PMC4496403.
137. Wang T, Qin ZY, Wen LZ, Guo Y, Liu Q, Lei ZJ, Pan W, Liu KJ, Wang XW, Lai SJ, Sun WJ, Wei YL, Liu L, Guo L, Chen YQ, Wang J, Xiao HL, Bian XW, Chen DF, Wang B. Epigenetic restriction of Hippo signaling by MORC2 underlies stemness of hepatocellular carcinoma cells. *Cell Death Differ.* 2018;25(12):2086-100. Epub 20180319. doi: 10.1038/s41418-018-0095-6. PubMed PMID: 29555977; PMCID: PMC6261965.
138. Mandal K. Review of PIP2 in Cellular Signaling, Functions and Diseases. *Int J Mol Sci.* 2020;21(21). Epub 20201106. doi: 10.3390/ijms21218342. PubMed PMID: 33172190; PMCID: PMC7664428.
139. Nitzsche A, Pietila R, Love DT, Testini C, Ninchoji T, Smith RO, Ekvaryn E, Larsson J, Roche FP, Egana I, Jauhainen S, Berger P, Claesson-Welsh L, Hellstrom M. Paladin is a phosphoinositide phosphatase regulating endosomal VEGFR2 signalling and angiogenesis. *EMBO Rep.* 2021;22(2):e50218. Epub 20201228. doi: 10.15252/embr.202050218. PubMed PMID: 33369848; PMCID: PMC7857541.
140. Jungmichel S, Sylvestersen KB, Choudhary C, Nguyen S, Mann M, Nielsen ML. Specificity and commonality of the phosphoinositide-binding proteome analyzed by quantitative mass spectrometry. *Cell Rep.* 2014;6(3):578-91. Epub 20140123. doi: 10.1016/j.celrep.2013.12.038. PubMed PMID: 24462288.
141. Durrant TN, Hutchinson JL, Heesom KJ, Anderson KE, Stephens LR, Hawkins PT, Marshall AJ, Moore SF, Hers I. In-depth PtdIns(3,4,5)P(3) signalosome analysis identifies DAPP1 as a negative regulator of GPVI-driven platelet function. *Blood Adv.* 2017;1(14):918-32. doi: 10.1182/bloodadvances.2017005173. PubMed PMID: 29242851; PMCID: PMC5726495.
142. Woo JH, Shimoni Y, Yang WS, Subramaniam P, Iyer A, Nicoletti P, Rodriguez Martinez M, Lopez G, Mattioli M, Realubit R, Karan C, Stockwell BR, Bansal M, Califano A. Elucidating Compound Mechanism of Action by Network Perturbation Analysis. *Cell.* 2015;162(2):441-51. doi: 10.1016/j.cell.2015.05.056. PubMed PMID: 26186195; PMCID: PMC4506491.
143. Tsai CW, Lai FJ, Sheu HM, Lin YS, Chang TH, Jan MS, Chen SM, Hsu PC, Huang TT, Huang TC, Sheen MC, Chen ST, Chang WC, Chang NS, Hsu LJ. WWOX suppresses autophagy for inducing apoptosis in methotrexate-treated human squamous cell carcinoma. *Cell Death Dis.* 2013;4(9):e792. Epub 20130905. doi: 10.1038/cddis.2013.308. PubMed PMID: 24008736; PMCID: PMC3789168.
144. Shin SB, Woo SU, Chin YW, Jang YJ, Yim H. Sensitivity of TP53-Mutated Cancer Cells to the Phytoestrogen Genistein Is Associated With Direct Inhibition of Plk1 Activity. *J Cell Physiol.* 2017;232(10):2818-28. Epub 20170410. doi: 10.1002/jcp.25680. PubMed PMID: 27861885.

145. Fontaine F, Overman J, Francois M. Pharmacological manipulation of transcription factor protein-protein interactions: opportunities and obstacles. *Cell Regen.* 2015;4(1):2. Epub 20150312. doi: 10.1186/s13619-015-0015-x. PubMed PMID: 25848531; PMCID: PMC4365538.
146. Li MM, Nilsen A, Shi Y, Fusser M, Ding YH, Fu Y, Liu B, Niu Y, Wu YS, Huang CM, Olofsson M, Jin KX, Lv Y, Xu XZ, He C, Dong MQ, Rendtlew Danielsen JM, Klungland A, Yang YG. ALKBH4-dependent demethylation of actin regulates actomyosin dynamics. *Nat Commun.* 2013;4:1832. doi: 10.1038/ncomms2863. PubMed PMID: 23673617; PMCID: PMC3674258.
147. Jingushi K, Aoki M, Ueda K, Kogaki T, Tanimoto M, Monoe Y, Ando M, Matsumoto T, Minami K, Ueda Y, Kitae K, Hase H, Nagata T, Harada-Takeda A, Yamamoto M, Kawahara K, Tabata K, Furukawa T, Sato M, Tsujikawa K. ALKBH4 promotes tumorigenesis with a poor prognosis in non-small-cell lung cancer. *Sci Rep.* 2021;11(1):8677. Epub 20210421. doi: 10.1038/s41598-021-87763-1. PubMed PMID: 33883577; PMCID: PMC8060266.
148. Xie L, Xie L, Bourne PE. Structure-based systems biology for analyzing off-target binding. *Curr Opin Struct Biol.* 2011;21(2):189-99. Epub 20110201. doi: 10.1016/j.sbi.2011.01.004. PubMed PMID: 21292475; PMCID: PMC3070778.
149. Metz JT, Hajduk PJ. Rational approaches to targeted polypharmacology: creating and navigating protein-ligand interaction networks. *Curr Opin Chem Biol.* 2010;14(4):498-504. Epub 20100706. doi: 10.1016/j.cbpa.2010.06.166. PubMed PMID: 20609615.
150. Sydow D, Assmann E, Kooistra AJ, Rippmann F, Volkamer A. KiSSim: Predicting Off-Targets from Structural Similarities in the Kinome. *J Chem Inf Model.* 2022;62(10):2600-16. Epub 20220510. doi: 10.1021/acs.jcim.2c00050. PubMed PMID: 35536589.
151. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. *Science.* 2002;298(5600):1912-34. doi: 10.1126/science.1075762. PubMed PMID: 12471243.
152. Franklin RA, McCubrey JA. Kinases: positive and negative regulators of apoptosis. *Leukemia.* 2000;14(12):2019-34. doi: 10.1038/sj.leu.2401967. PubMed PMID: 11187889.
153. Huang C, Jacobson K, Schaller MD. MAP kinases and cell migration. *J Cell Sci.* 2004;117(Pt 20):4619-28. doi: 10.1242/jcs.01481. PubMed PMID: 15371522.
154. Cohen P, Alessi DR. Kinase drug discovery--what's next in the field? *ACS Chem Biol.* 2013;8(1):96-104. Epub 20121231. doi: 10.1021/cb300610s. PubMed PMID: 23276252; PMCID: PMC4208300.
155. Cohen P, Cross D, Janne PA. Kinase drug discovery 20 years after imatinib: progress and future directions. *Nat Rev Drug Discov.* 2021;20(7):551-69. Epub 20210517. doi: 10.1038/s41573-021-00195-4. PubMed PMID: 34002056; PMCID: PMC8127496.
156. Noble ME, Endicott JA, Johnson LN. Protein kinase inhibitors: insights into drug design from structure. *Science.* 2004;303(5665):1800-5. doi: 10.1126/science.1095920. PubMed PMID: 15031492.
157. Lin A, Giuliano CJ, Palladino A, John KM, Abramowicz C, Yuan ML, Sausville EL, Lukow DA, Liu L, Chait AR, Galluzzo ZC, Tucker C, Sheltzer JM. Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Sci Transl Med.* 2019;11(509). doi: 10.1126/scitranslmed.aaw8412. PubMed PMID: 31511426; PMCID: PMC7717492.
158. Volkamer A, Eid S, Turk S, Rippmann F, Fulle S. Identification and Visualization of Kinase-Specific Subpockets. *J Chem Inf Model.* 2016;56(2):335-46. Epub 20160129. doi: 10.1021/acs.jcim.5b00627. PubMed PMID: 26735903.

159. Curran PR, Radoux CJ, Smilova MD, Sykes RA, Higuieruelo AP, Bradley AR, Marsden BD, Spring DR, Blundell TL, Leach AR, Pitt WR, Cole JC. Hotspots API: A Python Package for the Detection of Small Molecule Binding Hotspots and Application to Structure-Based Drug Design. *J Chem Inf Model.* 2020;60(4):1911-6. Epub 20200402. doi: 10.1021/acs.jcim.9b00996. PubMed PMID: 32207937.
160. Smilova MD, Curran PR, Radoux CJ, von Delft F, Cole JC, Bradley AR, Marsden BD. Fragment Hotspot Mapping to Identify Selectivity-Determining Regions between Related Proteins. *J Chem Inf Model.* 2022;62(2):284-94. Epub 20220112. doi: 10.1021/acs.jcim.1c00823. PubMed PMID: 35020376; PMCID: PMC8790751.
161. Chen BY, Honig B. VASP: a volumetric analysis of surface properties yields insights into protein-ligand binding specificity. *PLoS Comput Biol.* 2010;6(8). Epub 20100812. doi: 10.1371/journal.pcbi.1000881. PubMed PMID: 20814581; PMCID: PMC2930297.
162. Georgiev GD, Dodd KF, Chen BY. Precise parallel volumetric comparison of molecular surfaces and electrostatic isopotentials. *Algorithms Mol Biol.* 2020;15:11. Epub 20200525. doi: 10.1186/s13015-020-00168-z. PubMed PMID: 32489400; PMCID: PMC7247173.
163. Chen BY. VASP-E: specificity annotation with a volumetric analysis of electrostatic isopotentials. *PLoS Comput Biol.* 2014;10(8):e1003792. Epub 20140828. doi: 10.1371/journal.pcbi.1003792. PubMed PMID: 25166865; PMCID: PMC4148194.
164. Quintana FM, Kong Z, He L, Chen BY. DeepVASP-E: A Flexible Analysis of Electrostatic Isopotentials for Finding and Explaining Mechanisms that Control Binding Specificity. *Pac Symp Biocomput.* 2022;27:56-67. PubMed PMID: 34890136; PMCID: PMC9174418.
165. Chen BY, Bandyopadhyay S. A regionalizable statistical model of intersecting regions in protein-ligand binding cavities. *J Bioinform Comput Biol.* 2012;10(3):1242004. doi: 10.1142/S0219720012420048. PubMed PMID: 22809380.
166. Guo Z, Chen BY. Conformational Sampling Reveals Amino Acids with a Steric Influence on Specificity. *J Comput Biol.* 2015;22(9):861-75. doi: 10.1089/cmb.2015.0117. PubMed PMID: 26335806.
167. Lu H, Zhou Q, He J, Jiang Z, Peng C, Tong R, Shi J. Recent advances in the development of protein-protein interactions modulators: mechanisms and clinical trials. *Signal Transduct Target Ther.* 2020;5(1):213. Epub 20200923. doi: 10.1038/s41392-020-00315-3. PubMed PMID: 32968059; PMCID: PMC7511340.
168. Charlotiaux B, Zhong Q, Dreze M, Cusick ME, Hill DE, Vidal M. Protein-protein interactions and networks: forward and reverse edgetics. *Methods Mol Biol.* 2011;759:197-213. doi: 10.1007/978-1-61779-173-4\_12. PubMed PMID: 21863489.
169. Rosell M, Fernandez-Recio J. Hot-spot analysis for drug discovery targeting protein-protein interactions. *Expert Opin Drug Discov.* 2018;13(4):327-38. Epub 20180129. doi: 10.1080/17460441.2018.1430763. PubMed PMID: 29376444.
170. Olah J, Szenasi T, Lehotzky A, Norris V, Ovadi J. Challenges in Discovering Drugs That Target the Protein-Protein Interactions of Disordered Proteins. *Int J Mol Sci.* 2022;23(3). Epub 20220128. doi: 10.3390/ijms23031550. PubMed PMID: 35163473; PMCID: PMC8835748.
171. Smith MC, Gestwicki JE. Features of protein-protein interactions that translate into potent inhibitors: topology, surface area and affinity. *Expert Rev Mol Med.* 2012;14:e16. Epub 20120726. doi: 10.1017/erm.2012.10. PubMed PMID: 22831787; PMCID: PMC3591511.

172. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES. Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol.* 2007;25(1):71-5. doi: 10.1038/nbt1273. PubMed PMID: 17211405.
173. Buchwald P. Small-molecule protein-protein interaction inhibitors: therapeutic potential in light of molecular size, chemical space, and ligand binding efficiency considerations. *IUBMB Life.* 2010;62(10):724-31. doi: 10.1002/iub.383. PubMed PMID: 20979208.
174. Diaz-Eufracio BI, Naveja JJ, Medina-Franco JL. Protein-Protein Interaction Modulators for Epigenetic Therapies. *Adv Protein Chem Struct Biol.* 2018;110:65-84. Epub 20170725. doi: 10.1016/bs.apcsb.2017.06.002. PubMed PMID: 29413000.
175. Arkin MR, Wells JA. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat Rev Drug Discov.* 2004;3(4):301-17. doi: 10.1038/nrd1343. PubMed PMID: 15060526.
176. Ivanov AA, Khuri FR, Fu H. Targeting protein-protein interactions as an anticancer strategy. *Trends Pharmacol Sci.* 2013;34(7):393-400. Epub 20130529. doi: 10.1016/j.tips.2013.04.007. PubMed PMID: 23725674; PMCID: PMC3773978.
177. Souers AJ, Levenson JD, Boghaert ER, Ackler SL, Catron ND, Chen J, Dayton BD, Ding H, Enschede SH, Fairbrother WJ, Huang DC, Hymowitz SG, Jin S, Khaw SL, Kovar PJ, Lam LT, Lee J, Maecker HL, Marsh KC, Mason KD, Mitten MJ, Nimmer PM, Oleksijew A, Park CH, Park CM, Phillips DC, Roberts AW, Sampath D, Seymour JF, Smith ML, Sullivan GM, Tahir SK, Tse C, Wendt MD, Xiao Y, Xue JC, Zhang H, Humerickhouse RA, Rosenberg SH, Elmore SW. ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. *Nat Med.* 2013;19(2):202-8. Epub 20130106. doi: 10.1038/nm.3048. PubMed PMID: 23291630.
178. Lasica M, Anderson MA. Review of Venetoclax in CLL, AML and Multiple Myeloma. *J Pers Med.* 2021;11(6). Epub 20210524. doi: 10.3390/jpm11060463. PubMed PMID: 34073976; PMCID: PMC8225137.
179. Shin WH, Kumazawa K, Imai K, Hirokawa T, Kihara D. Current Challenges and Opportunities in Designing Protein-Protein Interaction Targeted Drugs. *Adv Appl Bioinform Chem.* 2020;13:11-25. Epub 20201112. doi: 10.2147/AABC.S235542. PubMed PMID: 33209039; PMCID: PMC7669531.
180. Basse MJ, Betzi S, Bourgeas R, Bouzidi S, Chetrit B, Hamon V, Morelli X, Roche P. 2P2Idb: a structural database dedicated to orthosteric modulation of protein-protein interactions. *Nucleic Acids Res.* 2013;41(Database issue):D824-7. Epub 20121130. doi: 10.1093/nar/gks1002. PubMed PMID: 23203891; PMCID: PMC3531195.
181. Basse MJ, Betzi S, Morelli X, Roche P. 2P2Idb v2: update of a structural database dedicated to orthosteric modulation of protein-protein interactions. *Database (Oxford).* 2016;2016. Epub 20160315. doi: 10.1093/database/baw007. PubMed PMID: 26980515; PMCID: PMC4792518.
182. Ikeda K, Maezawa Y, Yonezawa T, Shimizu Y, Tashiro T, Kanai S, Sugaya N, Masuda Y, Inoue N, Niimi T, Masuya K, Mizuguchi K, Furuya T, Osawa M. DLiP-PPI library: An integrated chemical database of small-to-medium-sized molecules targeting protein-protein interactions. *Front Chem.* 2022;10:1090643. Epub 20230109. doi: 10.3389/fchem.2022.1090643. PubMed PMID: 36700083; PMCID: PMC9868583.
183. Labbe CM, Kuenemann MA, Zarzycka B, Vriend G, Nicolaes GA, Lagorce D, Miteva MA, Villoutreix BO, Sperandio O. iPPI-DB: an online database of modulators of protein-protein

interactions. *Nucleic Acids Res.* 2016;44(D1):D542-7. Epub 20151001. doi: 10.1093/nar/gkv982. PubMed PMID: 26432833; PMCID: PMC4702945.

184. Da Silva F, Rognan D. Structure-Based Detection of Orthosteric and Allosteric Pockets at Protein-Protein Interfaces. *Methods Mol Biol.* 2018;1825:281-94. doi: 10.1007/978-1-4939-8639-2\_8. PubMed PMID: 30334209.

185. Skolnick J, Zhou H. Implications of the Essential Role of Small Molecule Ligand Binding Pockets in Protein-Protein Interactions. *J Phys Chem B.* 2022;126(36):6853-67. Epub 20220831. doi: 10.1021/acs.jpcc.2c04525. PubMed PMID: 36044742; PMCID: PMC9484464.

186. Guharoy M, Chakrabarti P. Conservation and relative importance of residues across protein-protein interfaces. *Proc Natl Acad Sci U S A.* 2005;102(43):15447-52. Epub 20051012. doi: 10.1073/pnas.0505425102. PubMed PMID: 16221766; PMCID: PMC1266102.

187. David A, Sternberg MJ. The Contribution of Missense Mutations in Core and Rim Residues of Protein-Protein Interfaces to Human Disease. *J Mol Biol.* 2015;427(17):2886-98. Epub 20150711. doi: 10.1016/j.jmb.2015.07.004. PubMed PMID: 26173036; PMCID: PMC4548493.

188. Johnson DK, Karanicolas J. Selectivity by small-molecule inhibitors of protein interactions can be driven by protein surface fluctuations. *PLoS Comput Biol.* 2015;11(2):e1004081. Epub 20150223. doi: 10.1371/journal.pcbi.1004081. PubMed PMID: 25706586; PMCID: PMC4338137.

189. Zhang QC, Deng L, Fisher M, Guan J, Honig B, Petrey D. PredUs: a web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Res.* 2011;39(Web Server issue):W283-7. Epub 20110523. doi: 10.1093/nar/gkr311. PubMed PMID: 21609948; PMCID: PMC3125747.

190. Zhang QC, Petrey D, Norel R, Honig BH. Protein interface conservation across structure space. *Proc Natl Acad Sci U S A.* 2010;107(24):10896-901. Epub 20100601. doi: 10.1073/pnas.1005894107. PubMed PMID: 20534496; PMCID: PMC2890749.

191. Guharoy M, Chakrabarti P. Conserved residue clusters at protein-protein interfaces and their use in binding site identification. *BMC Bioinformatics.* 2010;11:286. Epub 20100527. doi: 10.1186/1471-2105-11-286. PubMed PMID: 20507585; PMCID: PMC2894039.

192. Shulman-Peleg A, Shatsky M, Nussinov R, Wolfson HJ. Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biol.* 2007;5:43. Epub 20071009. doi: 10.1186/1741-7007-5-43. PubMed PMID: 17925020; PMCID: PMC2231411.

193. Da Silva F, Desaphy J, Rognan D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions. *ChemMedChem.* 2018;13(6):507-10. Epub 20171107. doi: 10.1002/cmdc.201700505. PubMed PMID: 29024463; PMCID: PMC5901026.

194. Desaphy J, Azdimousa K, Kellenberger E, Rognan D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *J Chem Inf Model.* 2012;52(8):2287-99. Epub 20120816. doi: 10.1021/ci300184x. PubMed PMID: 22834646.

195. Desaphy J, Bret G, Rognan D, Kellenberger E. sc-PDB: a 3D-database of ligandable binding sites--10 years on. *Nucleic Acids Res.* 2015;43(Database issue):D399-404. Epub 20141009. doi: 10.1093/nar/gku928. PubMed PMID: 25300483; PMCID: PMC4384012.

196. Kellenberger E, Muller P, Schalon C, Bret G, Foata N, Rognan D. sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J Chem Inf Model.* 2006;46(2):717-27. doi: 10.1021/ci050372x. PubMed PMID: 16563002.

197. Meslamani J, Rognan D, Kellenberger E. sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics*. 2011;27(9):1324-6. Epub 20110312. doi: 10.1093/bioinformatics/btr120. PubMed PMID: 21398668.
198. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. Pfam: The protein families database in 2021. *Nucleic Acids Res*. 2021;49(D1):D412-D9. doi: 10.1093/nar/gkaa913. PubMed PMID: 33125078; PMCID: PMC7779014.
199. Gao M, Skolnick J. APoc: large-scale identification of similar protein pockets. *Bioinformatics*. 2013;29(5):597-604. Epub 20130117. doi: 10.1093/bioinformatics/btt024. PubMed PMID: 23335017; PMCID: PMC3582269.
200. Lee HS, Im W. G-LoSA: An efficient computational tool for local structure-centric biological studies and drug design. *Protein Sci*. 2016;25(4):865-76. Epub 20160306. doi: 10.1002/pro.2890. PubMed PMID: 26813336; PMCID: PMC4941214.
201. von Behren MM, Volkamer A, Henzler AM, Schomburg KT, Urbaczek S, Rarey M. Fast protein binding site comparison via an index-based screening technology. *J Chem Inf Model*. 2013;53(2):411-22. Epub 20130207. doi: 10.1021/ci300469h. PubMed PMID: 23390978.
202. Batista J, Hawkins, P.C., Tolbert, R., Geballe, M.T. SiteHopper - a unique tool for binding site comparison *J Cheminform*. 2014;6.
203. Sael L, Kihara D. Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. *Proteins*. 2012;80(4):1177-95. Epub 20120124. doi: 10.1002/prot.24018. PubMed PMID: 22275074; PMCID: PMC3294165.
204. Bhadra A, Yeturu, K. Site2Vec: a reference frame invariant algorithm for vector embedding of protein–ligand binding sites. *Mach Learn: Sci Technol*. 2021;2.
205. Simonovsky M, Meyers J. DeeplyTough: Learning Structural Comparison of Protein Binding Sites. *J Chem Inf Model*. 2020;60(4):2356-66. Epub 20200318. doi: 10.1021/acs.jcim.9b00554. PubMed PMID: 32023053.
206. Xu Q, Dunbrack RL, Jr. Assignment of protein sequences to existing domain and family classification systems: Pfam and the PDB. *Bioinformatics*. 2012;28(21):2763-72. Epub 20120831. doi: 10.1093/bioinformatics/bts533. PubMed PMID: 22942020; PMCID: PMC3476341.
207. Bietz S, Urbaczek S, Schulz B, Rarey M. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J Cheminform*. 2014;6:12. Epub 20140403. doi: 10.1186/1758-2946-6-12. PubMed PMID: 24694216; PMCID: PMC4019353.
208. Fahrrolfes R, Bietz S, Flachsenberg F, Meyder A, Nittinger E, Otto T, Volkamer A, Rarey M. ProteinsPlus: a web portal for structure analysis of macromolecules. *Nucleic Acids Res*. 2017;45(W1):W337-W43. doi: 10.1093/nar/gkx333. PubMed PMID: 28472372; PMCID: PMC5570178.
209. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25(13):1605-12. doi: 10.1002/jcc.20084. PubMed PMID: 15264254.
210. Bargmann CI. Comparative chemosensation from receptors to ecology. *Nature*. 2006;444(7117):295-301. doi: 10.1038/nature05402. PubMed PMID: 17108953.
211. Bear DM, Lassance JM, Hoekstra HE, Datta SR. The Evolving Neural and Genetic Architecture of Vertebrate Olfaction. *Curr Biol*. 2016;26(20):R1039-R49. doi: 10.1016/j.cub.2016.09.011. PubMed PMID: 27780046; PMCID: PMC5104188.

212. Robertson HM. Molecular Evolution of the Major Arthropod Chemoreceptor Gene Families. *Annu Rev Entomol.* 2019;64:227-42. Epub 20181012. doi: 10.1146/annurev-ento-020117-043322. PubMed PMID: 30312552.
213. Del Marmol J, Yedlin MA, Ruta V. The structural basis of odorant recognition in insect olfactory receptors. *Nature.* 2021;597(7874):126-31. Epub 20210804. doi: 10.1038/s41586-021-03794-8. PubMed PMID: 34349260; PMCID: PMC8410599.
214. Buck L, Axel R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell.* 1991;65(1):175-87. doi: 10.1016/0092-8674(91)90418-x. PubMed PMID: 1840504.
215. Zhang X, Firestein S. The olfactory receptor gene superfamily of the mouse. *Nat Neurosci.* 2002;5(2):124-33. doi: 10.1038/nn800. PubMed PMID: 11802173.
216. Sato K, Pellegrino M, Nakagawa T, Nakagawa T, Vosshall LB, Touhara K. Insect olfactory receptors are heteromeric ligand-gated ion channels. *Nature.* 2008;452(7190):1002-6. Epub 20080413. doi: 10.1038/nature06850. PubMed PMID: 18408712.
217. Wicher D, Schafer R, Bauernfeind R, Stensmyr MC, Heller R, Heinemann SH, Hansson BS. *Drosophila* odorant receptors are both ligand-gated and cyclic-nucleotide-activated cation channels. *Nature.* 2008;452(7190):1007-11. Epub 20080413. doi: 10.1038/nature06861. PubMed PMID: 18408711.
218. Larsson MC, Domingos AI, Jones WD, Chiappe ME, Amrein H, Vosshall LB. Or83b encodes a broadly expressed odorant receptor essential for *Drosophila* olfaction. *Neuron.* 2004;43(5):703-14. doi: 10.1016/j.neuron.2004.08.019. PubMed PMID: 15339651.
219. Butterwick JA, Del Marmol J, Kim KH, Kahlson MA, Rogow JA, Walz T, Ruta V. Cryo-EM structure of the insect olfactory receptor Orco. *Nature.* 2018;560(7719):447-52. Epub 20180815. doi: 10.1038/s41586-018-0420-8. PubMed PMID: 30111839; PMCID: PMC6129982.
220. Barrio-Hernandez I, Yeo J, Janes J, Mirdita M, Gilchrist CLM, Wein T, Varadi M, Velankar S, Beltrao P, Steinegger M. Clustering predicted structures at the scale of the known protein universe. *Nature.* 2023;622(7983):637-45. Epub 20230913. doi: 10.1038/s41586-023-06510-w. PubMed PMID: 37704730; PMCID: PMC10584675.
221. Davis FP. Proteome-wide prediction of overlapping small molecule and protein binding sites using structure. *Mol Biosyst.* 2011;7(2):545-57. Epub 20101124. doi: 10.1039/c0mb00200c. PubMed PMID: 21103609.
222. Wishart DS, Guo A, Oler E, Wang F, Anjum A, Peters H, Dizon R, Sayeeda Z, Tian S, Lee BL, Berjanskii M, Mah R, Yamamoto M, Jovel J, Torres-Calzada C, Hiebert-Giesbrecht M, Lui VW, Varshavi D, Varshavi D, Allen D, Arndt D, Khetarpal N, Sivakumaran A, Harford K, Sanford S, Yee K, Cao X, Budinski Z, Liigand J, Zhang L, Zheng J, Mandal R, Karu N, Dambrova M, Schioth HB, Greiner R, Gautam V. HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res.* 2022;50(D1):D622-D31. doi: 10.1093/nar/gkab1062. PubMed PMID: 34986597; PMCID: PMC8728138.
223. Feng S, Park S, Choi YK, Im W. CHARMM-GUI Membrane Builder: Past, Current, and Future Developments and Applications. *J Chem Theory Comput.* 2023;19(8):2161-85. Epub 20230404. doi: 10.1021/acs.jctc.2c01246. PubMed PMID: 37014931; PMCID: PMC10174225.