

COMPUTATIONAL METHODS TO IDENTIFY AND TARGET DRUGGABLE BINDING
SITES AT PROTEIN-PROTEIN INTERACTIONS IN THE HUMAN PROTEOME

David Xu

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

September 2019

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Huanmei Wu, PhD, Co-Chair

Samy Meroueh, PhD, Co-Chair

April 15, 2019

Xiaowen Liu, PhD

Sarath Chandra Janga, PhD

Yunlong Liu, PhD

© 2019
David Xu

DEDICATION

To my family, without whom this journey would not be possible.

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere appreciation to my advisor Dr. Samy Meroueh for his support during my graduate studies and related research over the past years. I would also like to thank him for his insight, generosity, and support throughout this experience, without whom this would not be possible.

Sincere gratitude is given to my committee co-chair, Dr. Huanmei Wu, for her time and guidance through the Ph.D. process. I would also like to thank all my committee members, Dr. Xiaowen Liu, Dr. Sarath Chandra Janga, and Dr. Yunlong Liu, for their valuable comments and suggestions on my work.

I would like to express my sincerest gratitude to all my colleagues through the years in the Center for Computational Biology and Bioinformatics: Dr. Liwei Li, Dr. Yubing Si, Dr. Michael Wu, Dr. Jack Chiang, Dr. Hai Lin, Bo Wang, and many others, for thoughtful discussions and valuable insights into the interdisciplinary field of bioinformatics. I would also like to thank the other members of the lab, Dr. Khuchtumur Bum-Erdene, Dr. Donghui Zhou, Dr. Degang Liu, and Mona Ghozayel, without whom much of this work would remain computational and without validation. I am grateful to collaborators at the Open Science Grid, including Rob Quick, Scott Teige, and Mats Rynge, for providing high-throughput computational resources without which much of this work would not be possible. I am indebted to Dr. Yaoqi Zhou, who guided me through the first year of my MS study. I am thankful to my undergraduate research advisor, Dr. John Cheeseman, who introduced me to the research process.

Finally, I would like to express my deepest love to my family. My dearest thanks to my late father, Dr. Jigeng Xu, who passed away unexpected during my Ph.D. journey, but has forever instilled the drive for hard work and focus that made this work possible. To my mother, Li Lu, who taught me to be honest and warm, who I know will forever support me in my journey wherever it may go. To my little sister, Alice, for her love and support over the years.

David Xu

COMPUTATIONAL METHODS TO IDENTIFY AND TARGET DRUGGABLE BINDING
SITES AT PROTEIN-PROTEIN INTERACTIONS IN THE HUMAN PROTEOME

Protein-protein interactions are fundamental in cell signaling and cancer progression. An increasing prevalent idea in cancer therapy is the development of small molecules to disrupt protein-protein interactions. Small molecules impart their action by binding to pockets on the protein surface of their physiological target. At protein-protein interactions, these pockets are often too large and tight to be disrupted by conventional design techniques. Residues that contribute a disproportionate amount of energy at these interfaces are known as hot spots. The successful disruption of protein-protein interactions with small molecules is attributed to the ability of small molecules to mimic and engage these hot spots.

Here, the role of hot spots is explored in existing inhibitors and compared with the native protein ligand to explore how hot spot residues can be leveraged in protein-protein interactions. Few studies have explored the use of interface residues for the identification of hit compounds from structure-based virtual screening. The tight uPAR•uPA interaction offers a platform to test methods that leverage hot spots on both the protein receptor and ligand. A method is described that enriches for small molecules that both engage hot spots on the protein receptor uPAR and mimic hot spots on its protein ligand uPA. In addition, differences in chemical diversity in mimicking ligand hot spots is explored.

In addition to uPAR•uPA, there are additional opportunities at unperturbed protein-protein interactions implicated in cancer. Projects such as TCGA, which systematically catalog the hallmarks of cancer across multiple platforms, provide opportunities to identify novel protein-protein interactions that are paramount to cancer progression. To that end, a census of cancer-specific binding sites in the human proteome are identified to provide opportunities for drug discovery at the system level. Finally, tumor genomic, protein-protein interaction, and protein structural data is integrated to create chemogenomic libraries for phenotypic screening to uncover novel GBM targets and generate starting points for the development of GBM therapeutic agents.

Huanmei Wu, PhD, Co-Chair

Samy Meroueh, PhD, Co-Chair

TABLE OF CONTENTS

List of Tables	xii
List of Figures	xiii
Chapter 1. Introduction	1
1.1 Background	1
1.1.1 Cancer Genomics	1
1.1.2 Protein-Protein Interactions	1
1.2 Challenges Addressed	3
1.3 Major Contributions	4
Chapter 2. A Computational Investigation of Small-Molecule Engagement of Hot Spots at Protein-Protein Interaction Interfaces	11
2.1 Introduction	11
2.2 Results	12
2.2.1 Protein-Protein and Protein-Compound Complexes	12
2.2.2 Molecular Dynamics Simulations and Free Energy Calculations	18
2.2.3 Computational Alanine Scanning and Free Energy Decomposition	22
2.2.4 Bcl-xL•Bak	22
2.2.5 MDM2•p53	27
2.2.6 XIAP•Smac	30
2.2.7 IL-2•IL-2R α	34
2.2.8 BRD4•H4	38
2.2.9 Mimicking Hot Spots on the Protein Ligand	41
2.2.10 Effect of Native Protein Ligand and Small-Molecule Inhibitors on Receptor Dynamics	49
2.3 Discussion	49
2.4 Materials and Methods	51
2.4.1 Structural Preparation	51
2.4.2 Molecular Dynamics	52
2.4.3 Free Energy Calculations	53
2.4.4 Alanine Scanning	54
2.4.5 Decomposition Energy	54
2.4.6 Ligand Pharmacophore	56
2.4.7 Dynamic Cross-Correlation Matrix	56
2.4.8 Statistical Analysis	57

Chapter 3. Mimicking Intermolecular Interactions of Tight Protein-Protein Complexes for Small-Molecule Antagonists.....	58
3.1 Introduction	58
3.2 Results	59
3.2.1 uPAR•uPA as a Platform to Test Rank-Ordering Methods	59
3.2.2 A New Fingerprint Method to Rank-Order Compounds Based on their Ability to Mimic the Binding Profile of uPA to Residues on uPAR	62
3.2.3 Application of the Fingerprint Method to Rank-Order Compounds using uPAR Interface Residues	63
3.2.4 Selecting Rank-Ordered Compounds using uPA Interface Residues	66
3.2.5 Selecting Rank-Ordered Compounds using both uPA and uPAR Interface Residues	70
3.3 Discussion	72
3.4 Materials and Methods	82
3.4.1 Virtual Screening	82
3.4.2 uPAR Interface Residues	82
3.4.3 uPA Interface Residues.....	84
3.4.4 Selection of Compounds	84
3.4.5 Fluorescence Polarization (FP) Assay	85
3.4.6 Microtiter-Based ELISA for uPAR•uPA	85
3.4.7 Microscale Thermophoresis (MST).....	86
3.4.8 (E)-2-(4-mercaptostyryl)-1,3,3-trimethyl-3H-indol-1-ium (MSTI) Assay	86
3.4.9 Horseradish Peroxidase-Phenol Red (HRP-PR) Redox Activity Assay	86
3.4.10 High-Performance Liquid Chromatography-Mass Spectrometry (HPLC-MS).....	87
Chapter 4. Chemical Space Overlap with Critical Protein-Protein Interface Residues in Commercial and Specialized Small-Molecule Libraries.....	88
4.1 Introduction	88
4.2 Results	89
4.2.1 Analysis of Compound Collection Physicochemical Properties.....	89
4.2.2 Compound Overlap with Protein-Ligand Hot Spots at Protein-Protein Interaction Interfaces	95
4.2.3 uPAR•uPA	96
4.2.4 TEAD•YAP	102
4.2.5 Cav α •Cav β	107

4.2.6 Virtual Screening of Commercial Library Against Two Protein-Protein Interactions.....	112
4.3 Discussion	118
4.4 Materials and Methods	120
4.4.1 Ligand Preparation.....	120
4.4.2 Principal Component Analysis	120
4.4.3 Principal Moment of Inertia.....	121
4.4.4 Protein Preparation	121
4.4.5 Virtual Screening	121
4.4.6 Ligand Pharmacophore	121
4.4.7 Fluorescence Polarization (FP) Assay	122
Chapter 5. Small-Molecule Binding Sites to Explore Protein-Protein Interactions in the Cancer Proteome.....	124
5.1 Introduction	124
5.2 Results	126
5.2.1 Three-Dimensional Structures of Proteins Encoded by Differentially Expressed Genes.....	126
5.2.2 Identification of Binding Sites on Protein Structures at the PDB.....	127
5.2.3 Classification of Binding Sites.....	132
5.2.4 Cavities at Enzyme Active Sites	132
5.2.5 Cavities at Protein-Protein Interaction Interfaces	133
5.2.6 Proteins with Binding Sites Located at Both Enzyme Active Sites and Protein-Protein Interaction Interfaces	133
5.2.7 Unclassified Binding Sites.....	138
5.2.8 A Search of Protein-Protein Interaction Networks to Identify OTH Binding Sites Located at PPI Interfaces	141
5.2.9 Cancer Signaling Pathways.....	142
5.2.10 Correlation with Patient Survival for Proteins Encoded by Differentially Expressed Genes.....	144
5.2.11 Protein-Protein Interaction Network.....	147
5.2.12 New Unexplored Targets for the Development of Small-Molecule Probes and Cancer Therapeutics.....	147
5.2.13 Missense Mutations on Protein Structures.....	149
5.3 Discussion	150

5.4 Materials and Methods	158
5.4.1 Gene Expression	158
5.4.2 Protein Structures	158
5.4.3 Binding Site Identification	159
5.4.4 Binding Site Annotation	159
5.4.5 Survival Analysis	160
5.4.6 Signaling Pathway	160
5.4.7 Protein-Protein Interaction Network.....	160
5.4.8 Missense Mutations	160
Chapter 6. Tumor-Specific Chemogenomic Libraries by Structure-Based Enrichment for Glioblastoma Phenotypic Screening	162
6.1 Introduction	162
6.2 Results	164
6.2.1 Target Selection, Virtual Screening, and Rank-Ordering of Chemical Library	164
6.2.2 Exploring Compounds in Patient-Derived GBM Spheroids.....	174
6.2.3 Structure-Activity Relationship (SAR) of 1 (IPR-2025)	174
6.2.4 Additional Phenotypic Screens with Candidates with Fewer Predicted Targets	175
6.2.5 Compounds Inhibit Tube-Formation in Matrigel.....	181
6.2.6 Structural Analysis and RNA Sequencing to Uncover Compound 1 Mechanism of Action.....	181
6.2.7 Thermal Proteome Profiling to Identify Potential Targets of 1 (IPR-2025)	187
6.2.8 Integrated Analysis of Computational, RNA-Seq, and TPP Data for Potential Mechanisms of Action.....	187
6.2.9 Structural Analysis and RNA Sequencing to Uncover Compound 29 Mechanism of Action.....	192
6.3 Discussion	193
6.4 Materials and Methods	196
6.4.1 TCGA GBM Gene Expression and Somatic Mutations	196
6.4.2 Protein-Protein Interaction Network.....	196
6.4.3 Druggable Binding Sites	197
6.4.4 Virtual Screening Against GBM Network.....	197
6.4.5 Cell Culture.....	198

6.4.6 Three-Dimensional Culture Models	198
6.4.7 Invasion Assay	199
6.4.8 RNA-Seq of Compound-Treated Cells	199
6.4.9 Thermal Proteome Profiling	200
6.4.10 Gene Set Analysis	202
6.4.11 Cellular Thermal Shift Assay (CETSA)	202
Chapter 7. Summary	203
7.1 Conclusion.....	203
7.2 Suggested Future Work	205
Appendices.....	208
Appendix A. Reprint Permissions for Published Works	208
References.....	211
Curriculum Vitae	

LIST OF TABLES

Table 2.1. Characteristics of protein-protein interaction complexes	14
Table 2.2. Calculated free energies (\pm standard error) of protein-protein and protein-ligand complexes	16
Table 3.1. Profiles of analogs of compound 26 (IPR-2992)	75
Table 4.1. Parameters for principal component analysis for each of the eight input descriptors across the first three principal components	92
Table 4.2. Matching count for compounds in ChemDiv against hot spots on uPA	98
Table 4.3. Matching count for compounds in DOS against hot spots on uPA	99
Table 4.4. Matching count for compounds in SCUBIDOO against hot spots on uPA	100
Table 4.5. Matching count for compounds in ChemDiv against hot spots on Yap	104
Table 4.6. Matching count for compounds in DOS against hot spots on Yap	105
Table 4.7. Matching count for compounds in SCUBIDOO against hot spots on Yap	106
Table 4.8. Matching count for compounds in ChemDiv against hot spots on Cav α	109
Table 4.9. Matching count for compounds in DOS against hot spots on Cav α	110
Table 4.10. Matching count for compounds in SCUBIDOO against hot spots on Cav α	111
Table 5.1. Structural coverage of TCGA and the human proteome	128
Table 5.2. Distribution of protein structures and druggable binding sites among cancer types ($\log_2FC \geq 2.0$, $DS \geq 1.0$)	129
Table 5.3. Proteins with binding site that is both ENZ and PPI	135
Table 5.4. Proteins with both ENZ and PPI binding sites	137
Table 5.5. Proteins with potential PPI binding sites identified from search against PrePPI	140
Table 5.6. Mutations in binding site on overexpressed and clinically-relevant genes	153
Table 6.1. Synthesized derivatives of 1 (IPR-2025)	172
Table 6.2. Predicted protein targets of 1 (IPR-2025)	178
Table 6.3. Proteins with largest increase in melting temperature when treated with compound 1 ($\Delta T_m \geq 3$ °C)	184
Table 6.4. Predicted protein targets of 29 (IPR-196)	189

LIST OF FIGURES

Figure 1.1. Workflow for the investigation of the structural basis for protein-protein interactions.....	6
Figure 1.2. Workflow to rank-order compounds using protein receptor and ligand hot spots	7
Figure 1.3. Exploring the role of chemical libraries for protein-protein interaction inhibitors	8
Figure 1.4. Workflow used to identify cancer-specific druggable proteins	9
Figure 1.5. Screening the GBM-specific network using structure-based virtual screening.....	10
Figure 2.1. Comparison of free energies in protein-protein and protein-compound complexes....	20
Figure 2.2. Per-residue decomposition versus computational alanine scanning of protein-protein complexes	21
Figure 2.3. Bcl-xL•Bak	24
Figure 2.4. Bcl-xL•Bak comparison with inhibitors	25
Figure 2.5. MDM2•p53.....	28
Figure 2.6. MDM2•p53 comparison with inhibitors.....	29
Figure 2.7. XIAP•Smac	32
Figure 2.8. XIAP•Smac comparison with inhibitors.....	33
Figure 2.9. IL-2•IL-2R α	36
Figure 2.10. IL-2•IL-2R α comparison with inhibitors.....	37
Figure 2.11. BRD4•H4.....	39
Figure 2.12. BRD4•H4 comparison with inhibitors.....	40
Figure 2.13. Hot-spot residues on the protein ligand Bak and overlap with inhibitors in Bcl-xL•Bak	43
Figure 2.14. Hot-spot residues on the protein ligand p53 and overlap with inhibitors in MDM2•p53	44
Figure 2.15. Hot-spot residues on the protein ligand Smac and overlap with inhibitors in XIAP•Smac.....	45
Figure 2.16. Hot-spot residues on the protein ligand IL-2R α and overlap with inhibitors in IL-2•IL-2R α	46
Figure 2.17. Hot-spot residues on the protein ligand H4 and overlap with inhibitors in BRD4•H4	47
Figure 2.18. Similarity in the dynamics of inhibitors with the native ligand on the protein receptor	48
Figure 3.1. Structure of the uPAR•uPA binding pocket (PDB ID: 3BT1)	60

Figure 3.2. Workflow for the fingerprint method used to identify compounds that mimic the intermolecular binding interactions in the uPAR•uPA complex	61
Figure 3.3. A virtual screen utilizing the interface residues of uPAR and validation of hits.....	64
Figure 3.4. Screening the derivatives of compound 1 (IPR-2797)	65
Figure 3.5. A virtual screen utilizing four interface residues of uPA	67
Figure 3.6. Validation of hits of virtual screen utilizing four interface residues of uPA	68
Figure 3.7. A virtual screen utilizing interface residues on both uPAR and uPA.....	73
Figure 3.8. Testing the derivatives of 26 (IPR-2992) leads to 30 (IPR-3011).....	74
Figure 3.9. Concentration-dependents studies for 26 (IPR-2992) and 30 (IPR-3011)	77
Figure 3.10. Screening the derivatives of compound 28 (IPR-3089)	78
Figure 3.11. Screening the derivatives of compound 29 (IPR-3193)	79
Figure 4.1. Histograms of individual physicochemical properties of the ChemDiv, DOS, and SCUBIDOO compound collections	90
Figure 4.2. Principal component analysis (PCA) of physicochemical properties.....	91
Figure 4.3. Principal moments of inertia (PMI) illustrate the shape diversity of three compound collections	93
Figure 4.4. The protein complex of uPAR and uPA peptide	97
Figure 4.5. The protein complex of TEAD and YAP peptide	103
Figure 4.6. The protein complex of Ca _v β and Ca _v α	108
Figure 4.7. Virtual screening of commercial library against uPAR.....	114
Figure 4.8. Concentration-dependent of compounds identified from initial screening against uPAR	115
Figure 4.9. Binding modes of hit compounds.....	116
Figure 4.10. Distribution of docking scores (GlideScores) for all compounds in ChemDiv compared to the five active compounds identified	117
Figure 5.1. Examples of binding site annotations.....	130
Figure 5.2. Classification of enzyme types by EC codes	131
Figure 5.3. Examples of proteins with both ENZ and PPI binding sites	136
Figure 5.4. Examples of proteins with potentially allosteric OTH binding sites	139
Figure 5.5. Binding sites in cancer-related signaling pathways.....	143
Figure 5.6. Proteins with binding sites that are both overexpressed and correlate with patient outcome.....	145
Figure 5.7. Integrating druggable binding sites with protein-protein interaction networks outcome.....	146

Figure 5.8. Secondary structure composition of residues surrounding PPI binding sites	151
Figure 5.9. Proteins with missense mutations	152
Figure 5.10. Occurrence of individual missense mutations	154
Figure 6.1. Identification of druggable targets implicated in GBM	166
Figure 6.2. Differential expression analysis of 169 tumor and 5 normal GBM RNA-seq samples from TCGA	167
Figure 6.3. Protein-protein interaction subnetwork of GBM-specific targets	168
Figure 6.4. Rank-ordering of compounds by their predicted number of predicted GBM-specific targets	169
Figure 6.5. Screening compounds against GBM43 leads to the identification of 1 (IPR-2025)	170
Figure 6.6. Concentration-dependent studies of 1 (IPR-2025) against cancer models	171
Figure 6.7. Compound activities of resynthesized 1 and synthesized derivatives	173
Figure 6.8. Reducing the number of predicted targets to select compounds for exploration in patient-derived GBM spheroids	176
Figure 6.9. Concentration-dependent studies of additional hits against cancer models	177
Figure 6.10. RNA-Seq of 1 (IPR-2025) treated GBM43 cells	180
Figure 6.11. Thermal proteome profiling of 1 (IPR-2025) treated GBM43 cells	183
Figure 6.12. Comparison of proteins implicated by compound 1	185
Figure 6.13. Validation of RACK1 as a target of compound 1	186
Figure 6.14. RNA-Seq of 29 (IPR-196) treated GBM43 cells	190
Figure 6.15. Comparison of proteins implicated by compound 29	191

Chapter 1

INTRODUCTION

1.1 BACKGROUND

1.1.1 Cancer Genomics. Hallmarks of cancer are driven by perturbations in protein-protein interactions and signaling pathways [1]. Large-scale sequencing studies of human tumors such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) provide opportunities to uncover the genetic basis of the processes that drive cancer. Whole-genome gene expression profiling studies have been instrumental not only in classifying tumors and uncovering genetic alterations in cancer cells (mutations, copy number, and rearrangements), but as a rich source of potential targets in a variety of cancers [2-13]. These studies have been instrumental in identifying tumor subtypes and uncovering driver mutations of these diseases.

While TCGA has been successful in identified critical cancer driver alterations at the gene level, the contribution of genes to the dysregulation of oncogenic signaling pathways remain unclear. The characterization of the molecular landscapes of cancers have identified a subset of genes and their protein products that promote tumorigenesis and progression through driver mutations [14, 15]. These mutations result in tumor heterogeneity, whereby patients with similar cancers can exhibit different responses to traditional treatment options. Somatic mutations contribute to tumorigenesis by destabilizing protein structure and altering protein function [16]. Many of these altered proteins do not have prototypical enzyme active sites traditional seen in traditional single-target drug discovery efforts, and can only be targeted through their protein-protein interactions [17].

1.1.2 Protein-Protein Interactions. Protein-protein interactions (PPIs) control nearly every aspect of normal cellular function, including enzyme catalysis, DNA regulation, biological signaling, and immune response. These interactions also contribute to activating or suppressing signaling networks involved in pathological processes such as cancer [17, 18]. In cells, it is estimated that signaling pathways occur in a network of more than 200,000 protein-protein interactions [19-21]. Protein-protein interactions were previously considered undruggable, largely due to the size of the protein interfaces ($\sim 1000\text{-}2000 \text{ \AA}^2$) compared to traditional enzyme active sites ($\sim 300\text{-}500 \text{ \AA}^2$) [22, 23]. The protein interfaces are also generally flat and devoid of grooves and cavities present at traditional enzyme binding sites [24]. Unlike enzyme binding sites, there are no native small-molecule ligands or substrates that bind to the binding pocket and can act as a

natural starting point. These three factors represent significant challenges in current drug discovery efforts of protein-protein interactions despite their therapeutic importance.

However, not all protein-protein interactions share these limitations. Rather, protein-protein interactions range from transient to tight [25-27]. They have been classified as primary, secondary, or tertiary depending on the architecture at the interface of the complex [22]. Primary interfaces are generally simple, involving a short linear peptide bound to the surface of another protein. Secondary interactions consist of an α -helix or β -turn that is often ensconced into a well-defined cavity of the receptor. Tertiary interactions are more complex, sometimes involving multiple secondary structures such as α -helices and β -strands. The size of the contact surface increases from primary to tertiary, reaching more than 1,500 \AA^2 in some cases for tertiary interactions [28].

The interaction energy (ΔG) of protein-protein interactions is not evenly distributed across the entire interaction interface [29]. Hot spots are residues that contribute substantially to the protein-protein interaction. They can be located either on the protein ligand or on the receptor. Hot spots are generally identified by alanine scanning studies, where individual amino acids are mutated to alanine and the resulting impact on the binding affinity is measured using biochemical or biophysical methods [30]. Computational methods such as molecular dynamics simulations have also been successfully used [31, 32]. Through these studies, critical hot spot residues have been discovered on previously considered undruggable proteins [33, 34]. The amino acid composition of interfaces favor certain amino acids, such as hydrophobic aromatic residues like tyrosine or tryptophan [28, 35]. Charged residues such as arginine, lysine, and glutamic acid are also frequently found at interfaces and often engage residues through salt-bridge and π -cation interactions [36, 37]. As a result, these residues are often identified as hot spots. Hot spots are often assembled in tightly packed clusters on the interface and are often referred to as hot regions [38]. It has been suggested that the distribution of these regions play a role in how proteins can have multiple binding partners.

In this work, the current challenges in targeting protein-protein interactions at both a target-specific level and across the human interactome is explored. A series of studies describes my effort to first understand the structural basis of protein-protein interactions by comparing the engagement of existing inhibitors of protein-protein interactions with the native protein ligand at five protein-protein interactions that are critical to cancer cell signaling. We find that more potent inhibitors of these protein-protein interactions are better able to engage critical residues, or hot spots, located on the protein receptor as well as mimic hot spots present on the protein ligand. We then propose a method to represent the engagement of compounds from chemical libraries to a protein receptor as a bitwise fingerprint to rank-order compounds from computational screening. Cancer exhibits many

phenotypes, such as uncontrolled cell growth, invasion, and metastasis. We propose that to inhibit many of the phenotypes associated with tumorigenesis and tumor progression will require compounds that can target the underlying protein-protein interaction network associated with these phenotypes at more than one point. In the final part of this work, a structure-based approach is used to identify compounds that target genes that have been implicated in cancer progression. We first identify druggable binding sites across the cancer proteome. Then, we propose a method to enrich chemical libraries for phenotypic screening to identify compounds that can potentially target druggable binding sites on proteins implicated in cancer.

1.2 CHALLENGES ADDRESSED

While protein-protein interactions are implicated in the pathogenesis of diseases such as cancer, one challenge is the identification of potential druggable binding sites on these proteins. Despite the therapeutic potential of protein-protein interactions, current efforts in drug discovery are largely focused on targeting kinases, nuclear receptors, ion channels, and rhodopsin-like G protein-coupled receptors (GPCRs) [39]. The lack of favorable physicochemical properties makes protein-protein interactions unsuitable for drug discovery. However, recent advances in the design of protein-protein interaction inhibitors that target tight and stable interactions critical to cancer signaling have provided opportunities to identify critical intermolecular interactions between the native protein-ligand complex to enrich chemical libraries for potential inhibitors of these protein-protein interactions. In this work, we followed a structure-based approach to develop scoring approaches to identify compounds that can potentially inhibit protein-protein interactions. Often, one of the first steps in target-based drug discovery is the use of structure-based virtual screening to computationally enrich or rank-order a chemical library to a well-defined binding pocket on a target of interest. However, traditional scoring functions were developed for targeting binding sites on enzymes, and not suitable for protein-protein interactions. Here, we develop a scoring function that uses the native interaction as a guide to rank-order chemical libraries.

In addition to targeting individual protein-protein interactions implicated in cancer, we developed an approach to enrich chemical libraries for phenotypic screening. Rather than identifying compounds that target single proteins, phenotypic screening identifies compounds that modulate a specific tumor phenotype, for example, cell viability or cell invasion. Traditionally, this is done in a high-throughput manner, in which millions of compounds are screened. Often, compounds identified from this approach are non-specific to either the targets of interest and are toxic to both cancer and normal cells. In the second part of this work, we address this challenge by following a structure-based approach for cancer-specific drug discovery by (i) identifying relevant

cancer-specific targets with druggable binding sites, and (ii) developing methods to identify small compounds for cancer therapeutics.

1.3 MAJOR CONTRIBUTIONS

Here, four major projects are presented to address these challenges in identifying potential inhibitors of protein-protein interactions: (i) understanding the structural basis of protein-protein interactions, (ii) leveraging hot spots for the inhibition of individual protein-protein interactions, and (iii) identifying and targeting protein-protein interaction networks.

First, topics related to targeting individual protein-protein interactions are explored. In Chapter 2, the role of hot spots is explored by surveying existing protein-protein interaction inhibitors as summarized in **Fig. 1.1**. Computational methods are used to identify critical residues, also known as hot spots, at protein-protein interaction interfaces using alanine scanning and per-residue energy decomposition. Then, we explore engagement of compounds with receptor hot spots and investigated overlap between compounds and ligand hot spots.

Next, we leveraged this for structural-based virtual screening. Structural-based virtual screening is often the first step in target-based drug discovery. The goal of structural-based virtual screening is to enrich chemical libraries for potential compound candidates by docking large chemical libraries to a well-defined binding site on a target of interest. Molecular docking is used to predict the binding poses of each compound to the target, and scoring functions are used to evaluate these poses and rank-order compounds. However, traditional scoring functions are designed for well-defined binding sites on enzymes, nuclear receptors, G-protein coupled receptors, and other targets that have been well-studied in cancer and other diseases and are not suitable for protein-protein interactions. Therefore, we developed a method to exploit the interaction of small molecules with both receptor and ligand hot spots to identify potential leads of protein-protein interaction inhibitors at individual protein-protein interactions in Chapter 3. This scoring method was applied to the urokinase protein-protein interaction to rank-order commercial compounds, eventually leading to the discovery of a fragment-like compound that was optimized to inhibit the interaction with low single-digit micromolar binding affinity (**Fig. 1.2**).

In Chapter 4, the role of chemical diversity is explored in mimicking ligand hot spots by comparing three chemical libraries with different physicochemical properties and developmental ideologies (**Fig. 1.3**). Three different tight interactions for which hot-spot residues have been identified were selected for analysis in order to provide insight into how different areas of chemical space can be used for the discovery of potential inhibitors.

Cancer is a disease that affects the protein-protein interaction landscape. Disruptions in this landscape results in multiple phenotypes indicative of cancer, for example, uncontrolled cell growth and resistance of cell death. Often, these phenotypes are a result of disturbance of both previously discovered and still undiscovered interactions critical to tumorigenesis. There are additional opportunities at undiscovered protein-protein interactions implicated in cancer. Systematic cancer projects such as TCGA have identified novel targets critical to tumorigenesis and tumor progression. In Chapter 5, a set of cancer-specific druggable binding pockets in the human proteome are identified with respect to their putative function, role in cancer signaling pathways, and somatic mutations by integrating multiomic TCGA genomic data and structural data as summarized in **Fig. 1.4**. Finally, in Chapter 6, a method to enrich chemogenomic libraries for phenotypic screening using GBM-specific targets is described. In this approach, a large-scale protein-compound interaction matrix is generated from large-scale virtual screening of chemical libraries to targets implicated in GBM (**Fig. 1.5**). Then, compounds are rank ordered based on their ability to maximally target druggable binding sites on proteins implicated in the disease. The discovery of a compound that inhibits GBM phenotypes without affecting normal cell viability suggests that our approach to create tumor-specific chemogenomic libraries may hold promise for developing more efficacious treatments for incurable diseases like GBM.

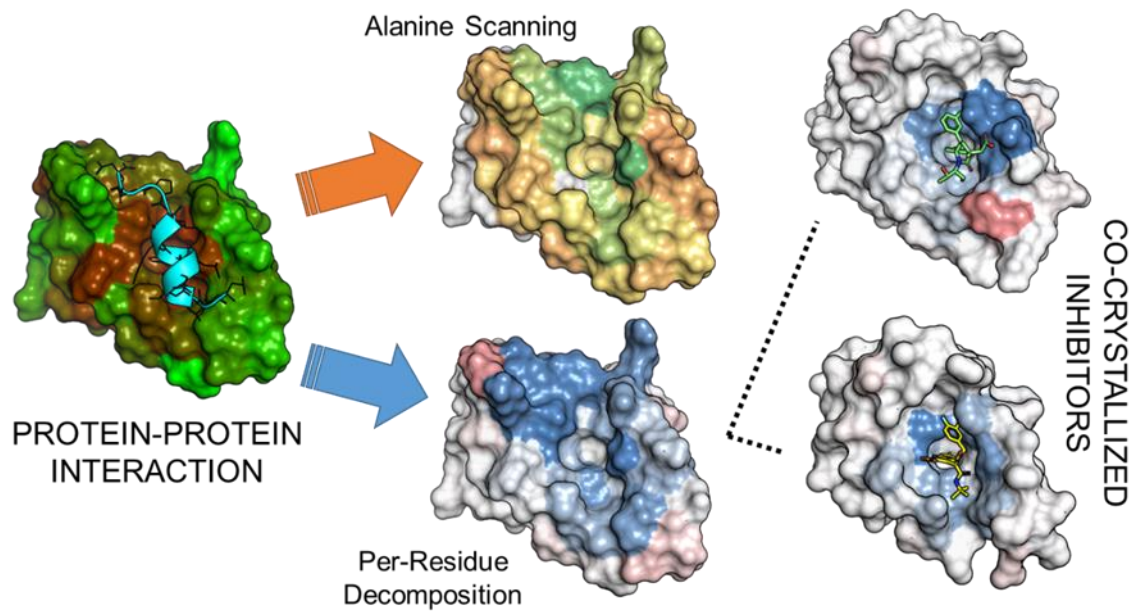


Figure 1.1. Workflow for the investigation of the structural basis for protein-protein interactions.

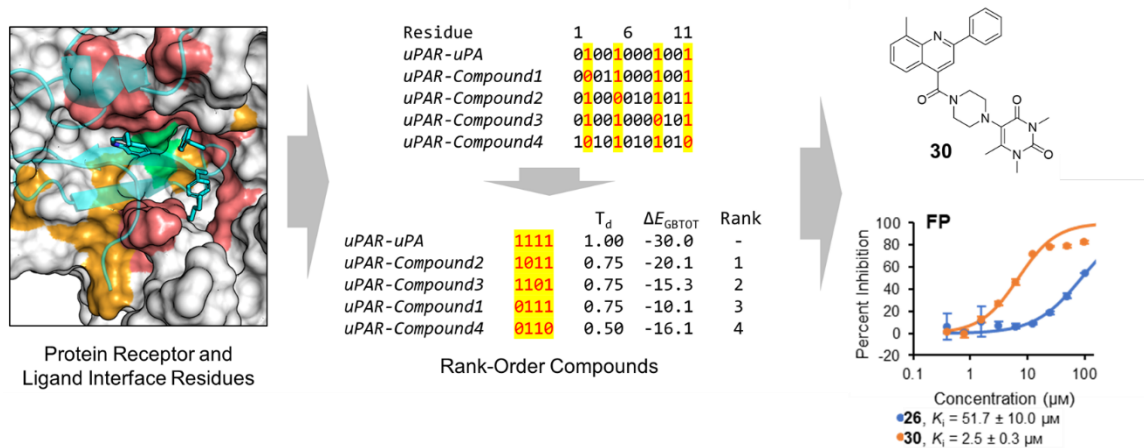


Figure 1.2. Workflow to rank-order compounds using protein receptor and ligand hot spots.

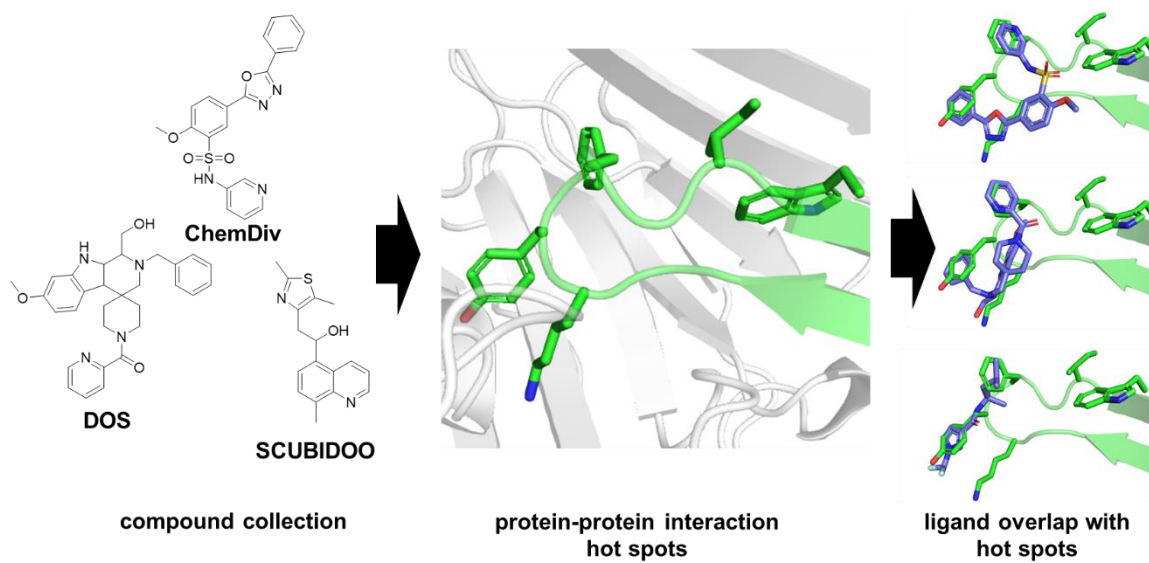


Figure 1.3. Exploring the role of chemical libraries for protein-protein interaction inhibitors.

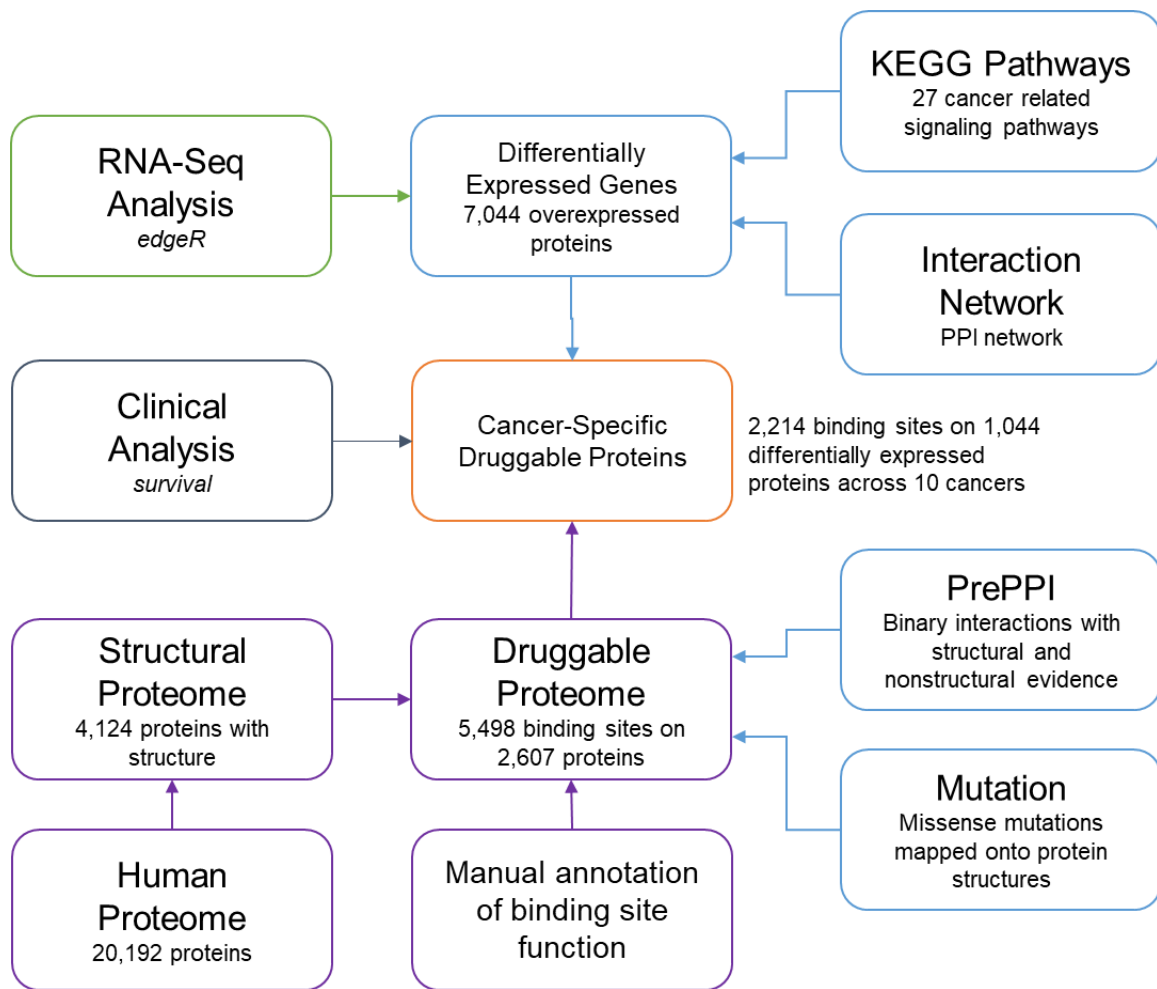


Figure 1.4. Workflow used to identify cancer-specific druggable proteins.

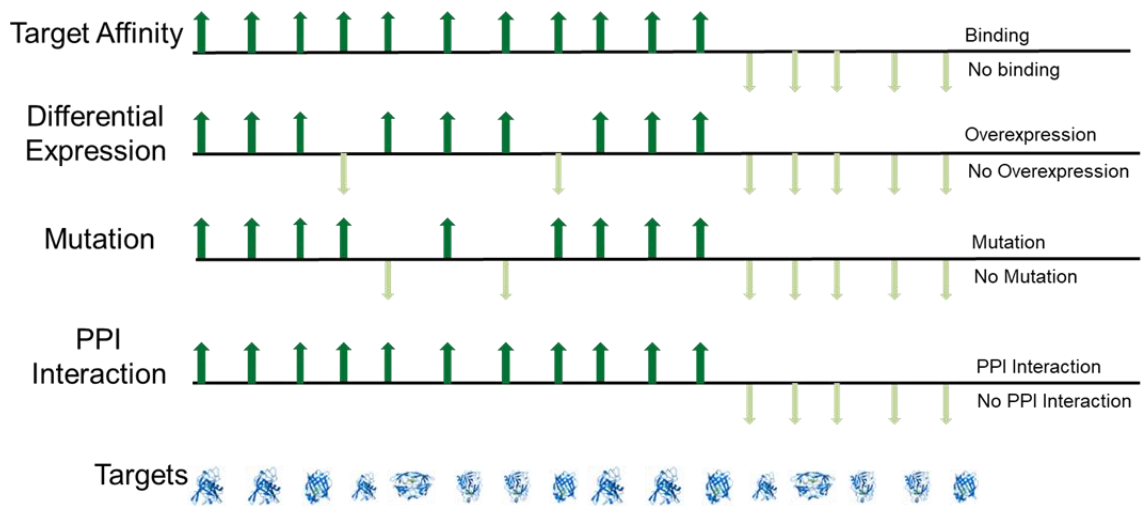
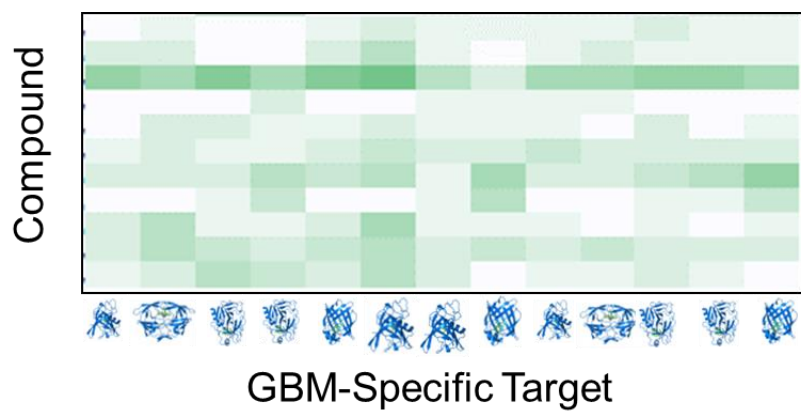


Figure 1.5. Screening the GBM-specific network using structure-based virtual screening.

Chapter 2

A COMPUTATIONAL INVESTIGATION OF SMALL-MOLECULE ENGAGEMENT OF HOT SPOTS AT PROTEIN-PROTEIN INTERACTION INTERFACES

2.1 INTRODUCTION

Small molecules disrupt tight protein-protein interactions by engaging or mimicking hot spots located at the protein-protein interface [28, 33, 38, 40, 41]. Hot spots are amino acids that contribute substantially to the protein-protein interaction. They can be located either on the protein ligand or on the receptor. Hot spots are generally identified by alanine scanning studies, where individual amino acids are mutated to alanine and the resulting impact on the binding affinity is measured using biochemical or biophysical methods [30]. Computational methods such as molecular dynamics simulations have also been successfully used [31, 32].

There is intense interest in the development of small organic molecules to disrupt protein-protein interactions [42]. Small molecules provide useful tools to dissect individual interactions of the cellular protein-protein interaction network. In addition, small molecules that disrupt protein-protein interactions associated with a disease can be further developed into therapeutic agents. Early strategies for developing protein-protein interaction inhibitors consisted of designing compounds with substituents that mimicked side chains of the protein ligand [43]. This has worked particularly well for the development of peptidomimetic inhibitors of protein-protein interaction [44] such as the MDM2•p53 interaction [45]. Another approach consists of searching for fragment-like compounds that bind to cavities at the protein-protein interface [46]. This method has led to nanomolar and sub-micromolar inhibitors of Bcl-xL•Bak [47], IL-2•IL-2R α [48], and more recently KEAP1•NRF2 [49].

Structure-based computational screening of commercially available chemical libraries has also been applied towards the discovery of small-molecule protein-protein interaction inhibitors. Virtual screening led to the discovery of fragment-like compounds that disrupted the interaction between IFN- α and its binding partner IFNAR [50]. Another strategy combining docking and pharmacophore definitions led to inhibitors of the LEDGF•p75 interaction [51]. It has been suggested that small molecules disrupt tight protein-protein interactions by engaging or mimicking hot spots located at the protein-protein interface [28, 33, 38, 40, 41, 43]. Despite the widely accepted view that disruption of hot-spot interactions is critical for the successful inhibition of protein-protein interactions, there is no systematic approach to take advantage of hot spots for the rational design of small-molecule antagonists.

Strategies that have designed compounds to mimic hot spots on the protein ligand generally ignore hot spots located on the protein receptor. Similarly, compounds that are designed using fragment-based methods are conceived to bind to pockets on the receptor protein without regard to hot spots located on the protein ligand. Understanding how compounds engage and mimic hot spots could help guide the design of chemical libraries and to guide structure-based computational screening of these chemical libraries for the discovery of small-molecule protein-protein interaction inhibitors.

Here, we subject protein-compound and protein-protein structures to explicit-solvent molecular dynamics simulations and free energy calculations. We select five protein-protein interactions that have been successfully inhibited with small molecules and for which there exists quality binding affinity data and co-crystal structures: Bcl-xL•Bak, MDM2•p53, XIAP•Smac, IL-2•IL-2R α , and BRD4•H4. For each protein-protein and protein-compound complex, MM-GBSA free energy calculations were carried out to determine the binding free energy for comparison to experimental binding affinities and IC₅₀s. In addition, for each protein-protein complex, we determine the free energy change due to mutation of interface residues to alanine (computational alanine scan). We explore the interaction of each compound and protein ligand to the predicted hot spots on the receptor using per-residue decomposition energy calculations. Furthermore, we use pharmacophore modeling to investigate how effectively compounds mimic hot spots located on the protein ligand. Finally, molecular dynamics simulations are analyzed to compare the effect of compounds on the dynamics of the receptor to those of the protein ligand.

2.2 RESULTS

2.2.1 Protein-Protein and Protein-Compound Complexes. Five protein-protein interactions that have been successfully inhibited previously with small molecules were selected for this work (**Table 2.1**). Two of these interactions are classified as primary, corresponding to a short linear peptide binding to a receptor protein: The Bir3 domain of X-linked inhibitor of apoptosis protein with a short peptide of Smac/DIABLO (XIAP•Smac; $K_d = 420 \pm 20$ nM), and the first of two bromodomains on BRD4 with a di-acetylated peptide from a histone 4 tail (BRD4•H4; $K_d = 4.8 \pm 0.4$ μ M). Another two interactions are classified as secondary: MDM2, an inhibitor of p53 transcriptional activation, with the tumor suppressor p53 (MDM2•p53; $K_d = 295$ nM), and a pro-survival protein Bcl-xL with a pro-apoptotic peptide of Bak (Bcl-xL•Bak; $K_d = 340 \pm 30$ nM). These consist of 13- and 16-residue α -helices bound to well-defined pockets on MDM2 and Bcl-xL, respectively. The last interaction is classified as tertiary: The cytokine interleukin-2 with its α -subunit (IL-2•IL-2R α ; $K_d = 13$ nM).

A total of 36 small-molecule inhibitors that were co-crystallized with their respective targets were considered (**Table 2.1**). The compounds have a wide range of chemical structures and physicochemical properties. The binding affinity of the compounds ranged from sub-nanomolar to sub-millimolar. The Bcl-xL compounds were generally the largest and exhibited the highest affinities and inhibition potency, with the majority showing nanomolar K_d and IC_{50} . The weakest affinity compound, **5**, was a lead compound that ultimately led to a sub-nanomolar inhibitor of Bcl-xL [52]. The inhibitors of the MDM2•p53 interaction had binding affinities to MDM2 that ranged from 0.4 to 916 nM, with similar IC_{50} values ranging from 1.1 to 1710 nM. Except for two of the seven XIAP antagonists, K_d s and IC_{50} s of these compounds were in the sub-micromolar range. Compounds **18** and **22** showed micromolar K_d . The IC_{50} for the four IL-2R antagonists ranged from 60 nM to 6000 nM. Finally, the BRD4 antagonists exhibited binding affinities that ranged from 36 nM to 2400 nM. The IC_{50} values ranged from 16 to 100 nM.

Ligand efficiency is defined as a compound's free energy of binding divided by the number of non-hydrogen atoms [53]. Generally, ligand efficiency for drug-like compounds should be greater than 0.30 [54]. Among the compounds that we have considered, only BRD4•H4 inhibitors exhibited ligand efficiencies that are greater than 0.3 (0.32 ± 0.01), while inhibitors of MDM2•p53 are below 0.3 with a ligand efficiency of 0.27 ± 0.02 . Inhibitors of XIAP•Smac have ligand efficiencies of 0.24 ± 0.02 , despite having similar number of heavy atoms as inhibitors of MDM2•p53 (XIAP•Smac: 35 ± 1 , MDM2•p53: 36 ± 2 , Mann-Whitney rank-sum test, $p = 0.88$). There are at least two halogen atoms in each of the MDM2•p53 antagonists compared to none in XIAP•Smac antagonists, resulting in approximately 80 Da increase in molecular weight despite the almost equal number of heavy atoms. Finally, Bcl-xL•Bak and IL-2•IL-2R α have ligand efficiencies of 0.22 ± 0.03 and 0.21 ± 0.01 , respectively. Overall, except for BRD4, most small-molecule protein-protein interaction inhibitors have poor ligand efficiencies.

The lipophilic efficiency of a compound measures the difference between its activity and lipophilicity [55]. Compounds with high lipophilicity tend to have increased target promiscuity and decreased solubility [55]. Thus, interactions involving compounds with high lipophilic efficiency are primarily directed and specific to a protein receptor [56]. Generally, lipophilic efficiency of lead-like compounds is greater than 5 [55]. Mean lipophilic efficiency across all inhibitors is 3.3 ± 0.3 , ranging from 1.9 ± 0.5 in MDM2•p53 to 5.7 ± 0.5 in XIAP•Smac. Among all the antagonists we have considered, only small-molecule XIAP•Smac antagonists and **26**, which inhibits IL-2•IL-2R α , have lipophilic efficiencies above 5.

Table 2.1. Characteristics of protein-protein interaction complexes.

PDB	Compound	Ligand	LE ^a	LLE ^b	K_d/K_i (nM)	IC ₅₀ (nM)	Ref
Bcl-xL•Bak							
1BXL		-				340 ± 30	[57]
1YSI	1	N3B	0.27	1.6	36 ± 1.6		[58]
2YXJ	2	N3C	0.23	4.4	0.37	2.5 ± 0.6	[59]
3QKD	3	HI0	0.20	3.1	4.2	3	[60]
3SP7	4	03B	0.19	2.1	<1	6 ± 1	[52]
3SPF	5	B50	0.14	0.6	138000 ± 76000	453000 ± 25000	[61]
4QVX	6	3CQ	0.32	6.2	<0.01		[62]
MDM2•p53							
1YCR		-			295		[63]
1RV1	7	IMZ	0.23	0.0		140	[64]
1T4E	8	DIZ	0.28	2.5	80	220	[65, 66]
3JZK	9	YIN	0.26	0.5		1230 ± 820	[67]
3LBK	10	K23	0.25	0.1	916	1710 ± 1103	[68]
3TU1	11	07G	0.27	3.1	250		[69]
3W69	12	LTZ	0.21	3.0		58	[70]
4DIJ	13	BLF	0.26	1.7		30	[71]
4ERE	14	OR2	0.33	3.2		4.2 ± 0.9	[72]
4ERF	15	OR3	0.38	4.3	0.4	1.1 ± 0.5	[72]
4HG7	16	NUT	0.24	0.7		71 ± 11	[73]
XIAP•Smac							
1G73		-			420 ± 20		[74]
2JK7	17	BI6	0.27	4.9	67 ± 18		[75]
2OPY	18	CO9	0.19	6.1	30000 ± 12000		[76]
3CLX	19	X22	0.25	5.9	250	270 ± 20	[77]
3CM2	20	X23	0.23	6.6	870	970 ± 120	[77]
3EYL	21	SMK	0.25	6.8	220	250 ± 40	[77, 78]
3HL5	22	9JZ	0.16	3.4	34000		[79]
5C83	23	4YN	0.30	5.5		160	[80]
IL-2•IL-2Rα							
1Z92		-			13		[81]
1M48	24	FRG	0.23	3.9	8200	3000	[81, 82]
1PW6	25	FRB	0.20	3.4		6000	[83]
1PY2	26	FRH	0.22	5.9		60	[83]
1QVN	27	FRI	0.19	2.8		250	[84]
BRD4•H4							
3UVW		-			4800 ± 400		[85]
2YEL	28	WSH	0.31	2.3	52.5	15.5 ± 1.9	[86]

3MXF	29	JQ1	0.32	2.5	49	77	[87, 88]
3P5O	30	EAM	0.33	3.9	55.2	36.1	[88]
3U5J	31	08H	0.35	2.0	2460 ± 110		[89]
3U5L	32	08K	0.37	2.1	640 ± 30		[89]
3ZYU	33	1GH	0.31	3.3		100	[90]
4F3I	34	0S6	0.36	3.6	36.1 ± 7.8	30 ± 4	[91]
4MR4	35	1K0	0.30	2.9	1142 ± 46		[92]
5D3L	36	57F	0.28	3.6	880		[93]

^a Ligand Efficiency, T = 298.15 K

^b Lipophilic Efficiency

Table 2.2 Calculated free energies (\pm standard error) of protein-protein and protein-ligand complexes.

PDB	Cpd	ΔE_{VDW}	ΔE_{ELE}	ΔE_{GB}	ΔE_{SURF}	ΔE_{GBTOT}	$\Delta G_{MM-GBSA}$
Bcl-xL•Bak							
1BXL		-93.5 \pm 0.4	-266.1 \pm 2.2	295.1 \pm 2.0	-13.6 \pm 0.0	-78.1 \pm 0.4	-36.7 \pm 0.5
1YSI	1	-49.4 \pm 0.1	-6.2 \pm 0.1	28.1 \pm 0.1	-6.3 \pm 0.0	-33.8 \pm 0.1	-16.7 \pm 0.3
2YXJ	2	-72.9 \pm 0.1	-433.4 \pm 0.9	458.9 \pm 0.9	-9.0 \pm 0.0	-56.3 \pm 0.1	-28.7 \pm 0.3
3QKD	3	-71.2 \pm 0.1	-384.9 \pm 0.7	406.1 \pm 0.7	-8.7 \pm 0.0	-58.6 \pm 0.1	-33.2 \pm 0.2
3SP7	4	-83.3 \pm 0.2	-223.2 \pm 0.7	238.8 \pm 0.6	-11.1 \pm 0.0	-78.8 \pm 0.2	-44.9 \pm 0.3
3SPF	5	-39.8 \pm 0.2	-5.1 \pm 0.3	20.3 \pm 0.2	-4.9 \pm 0.0	-29.5 \pm 0.2	-11.3 \pm 0.3
4QVX	6	-84.2 \pm 0.1	-181.0 \pm 0.5	199.6 \pm 0.5	-9.8 \pm 0.0	-75.4 \pm 0.1	-49.0 \pm 0.2
MDM2•p53							
1YCR		-73.5 \pm 0.2	-377.5 \pm 1.2	400.5 \pm 1.1	-10.1 \pm 0.0	-60.6 \pm 0.2	-25.1 \pm 0.3
1RV1	7	-44.4 \pm 0.1	-8.1 \pm 0.1	20.9 \pm 0.1	-5.1 \pm 0.0	-36.7 \pm 0.1	-17.8 \pm 0.2
1T4E	8	-43.8 \pm 0.1	-136.9 \pm 0.9	151.8 \pm 0.9	-5.1 \pm 0.0	-34.1 \pm 0.1	-15.9 \pm 0.2
3JZK	9	-40.3 \pm 0.1	-11.1 \pm 0.1	22.9 \pm 0.1	-4.5 \pm 0.0	-32.9 \pm 0.1	-14.3 \pm 0.2
3LBK	10	-36.6 \pm 0.1	-115.0 \pm 0.4	129.4 \pm 0.4	-4.5 \pm 0.0	-26.7 \pm 0.1	-9.7 \pm 0.2
3TU1	11	-42.4 \pm 0.1	-107.7 \pm 0.4	123.6 \pm 0.4	-5.4 \pm 0.0	-32.0 \pm 0.1	-12.9 \pm 0.2
3W69	12	-48.4 \pm 0.1	59.2 \pm 0.2	-44.6 \pm 0.2	-5.8 \pm 0.0	-39.7 \pm 0.1	-20.0 \pm 0.3
4DIJ	13	-44.7 \pm 0.1	-7.6 \pm 0.1	21.9 \pm 0.1	-5.3 \pm 0.0	-35.6 \pm 0.1	-16.1 \pm 0.2
4ERE	14	-40.3 \pm 0.1	-112.3 \pm 1.1	125.8 \pm 1.0	-5.0 \pm 0.0	-31.8 \pm 0.1	-12.0 \pm 0.2
4ERF	15	-38.7 \pm 0.1	-145.5 \pm 0.7	156.1 \pm 0.6	-5.2 \pm 0.0	-33.3 \pm 0.1	-13.7 \pm 0.2
4HG7	16	-45.4 \pm 0.1	-8.0 \pm 0.1	22.6 \pm 0.1	-5.4 \pm 0.0	-36.3 \pm 0.1	-16.6 \pm 0.3
XIAP•Smac							
1G73		-49.6 \pm 0.4	-265.8 \pm 2.4	280.6 \pm 2.3	-7.2 \pm 0.0	-42.0 \pm 0.3	-9.7 \pm 0.4
2JK7	17	-41.1 \pm 0.1	-133.0 \pm 0.5	131.4 \pm 0.4	-5.0 \pm 0.0	-47.8 \pm 0.1	-30.1 \pm 0.2
2OPY	18	-32.7 \pm 0.1	-162.3 \pm 1.1	171.2 \pm 1.0	-4.3 \pm 0.0	-28.1 \pm 0.2	-8.3 \pm 0.2
3CLX	19	-39.6 \pm 0.1	-178.7 \pm 0.5	178.1 \pm 0.4	-5.2 \pm 0.0	-45.4 \pm 0.1	-23.8 \pm 0.2
3CM2	20	-37.0 \pm 0.1	-271.4 \pm 0.8	266.2 \pm 0.7	-5.0 \pm 0.0	-47.2 \pm 0.1	-26.7 \pm 0.2
3EYL	21	-40.1 \pm 0.1	-247.5 \pm 0.7	240.3 \pm 0.6	-5.5 \pm 0.0	-52.6 \pm 0.2	-31.0 \pm 0.2
3HL5	22	-32.8 \pm 0.1	-146.2 \pm 0.4	142.2 \pm 0.3	-4.1 \pm 0.0	-40.9 \pm 0.1	-19.5 \pm 0.2
5C83	23	-43.4 \pm 0.1	-152.2 \pm 0.5	154.7 \pm 0.5	-4.9 \pm 0.0	-45.8 \pm 0.1	-23.7 \pm 0.3
IL-2•IL-2Ra							
1Z92		-73.3 \pm 0.8	-674.0 \pm 7.1	683.9 \pm 7.2	-12.9 \pm 0.1	-76.4 \pm 0.9	-28.9 \pm 0.9
1M48	24	-43.4 \pm 0.1	-129.7 \pm 0.4	129.5 \pm 0.3	-5.9 \pm 0.0	-49.4 \pm 0.1	-28.5 \pm 0.2
1PW6	25	-39.1 \pm 0.2	-118.9 \pm 0.4	120.6 \pm 0.3	-5.3 \pm 0.0	-42.8 \pm 0.2	-21.5 \pm 0.3
1PY2	26	-51.4 \pm 0.1	-228.9 \pm 0.8	234.0 \pm 0.8	-6.8 \pm 0.0	-53.1 \pm 0.1	-26.1 \pm 0.3
1QVN	27	-52.4 \pm 0.2	-133.7 \pm 0.4	140.3 \pm 0.4	-6.8 \pm 0.0	-52.5 \pm 0.1	-24.7 \pm 0.3
BRD4•H4							
3UVW		-70.8 \pm 0.2	-159.8 \pm 0.9	177.7 \pm 0.8	-10.6 \pm 0.0	-63.5 \pm 0.2	-31.3 \pm 0.4
2YEL	28	-40.7 \pm 0.1	-8.0 \pm 0.1	20.0 \pm 0.1	-5.0 \pm 0.0	-33.6 \pm 0.1	-16.1 \pm 0.2
3MXF	29	-41.4 \pm 0.1	-2.3 \pm 0.1	13.8 \pm 0.1	-4.9 \pm 0.0	-34.8 \pm 0.1	-17.7 \pm 0.2

3P5O	30	-41.5 ± 0.1	-9.9 ± 0.1	23.0 ± 0.1	-5.0 ± 0.0	-33.4 ± 0.1	-16.5 ± 0.2
3U5J	31	-34.9 ± 0.1	-9.8 ± 0.1	18.5 ± 0.1	-4.1 ± 0.0	-30.4 ± 0.1	-15.4 ± 0.2
3U5L	32	-36.0 ± 0.1	-11.7 ± 0.1	20.3 ± 0.1	-4.3 ± 0.0	-31.6 ± 0.1	-14.9 ± 0.2
3ZYU	33	-40.7 ± 0.1	-9.3 ± 0.1	22.4 ± 0.1	-4.9 ± 0.0	-32.4 ± 0.1	-15.5 ± 0.2
4F3I	34	-40.5 ± 0.1	-4.0 ± 0.1	15.0 ± 0.1	-4.9 ± 0.0	-34.3 ± 0.1	-17.6 ± 0.2
4MR4	35	-31.9 ± 0.1	-19.8 ± 0.2	28.6 ± 0.2	-4.4 ± 0.0	-27.6 ± 0.1	-11.7 ± 0.2
5D3L	36	-36.3 ± 0.1	-29.1 ± 0.3	39.8 ± 0.2	-4.9 ± 0.0	-30.5 ± 0.1	-13.4 ± 0.2

2.2.2 Molecular Dynamics Simulations and Free Energy Calculations. Molecular dynamics simulations and MM-GBSA calculations were carried out for the five protein-protein and 36 protein-compound complexes. The MM-GBSA free energies for individual complexes are reported in **Table 2.2**. The van der Waals potential energy (ΔE_{VDW}) and the free energy due to burial of solvent-accessible surface area (ΔE_{SURF}) were more favorable for the protein-protein complexes than the protein-compound complexes. The van der Waals potential energy ranged from -93.5 ± 0.4 kcal \cdot mol $^{-1}$ for Bcl-xL \cdot Bak to -49.6 ± 0.4 kcal \cdot mol $^{-1}$ for XIAP \cdot Smac, while ΔE_{SURF} ranged from -13.6 for Bcl-xL \cdot Bak to -7.2 kcal \cdot mol $^{-1}$ for XIAP \cdot Smac. ΔE_{SURF} is directly proportional to the change in solvent-accessible surface area upon binding [94]. The buried surface area on the receptor in the protein-protein complexes are approximately 940, 900, 715, 660, and 250 Å 2 for Bcl-xL, IL-2, BRD4, MDM2, and XIAP, respectively. Therefore, it was not a surprise to find that ΔE_{SURF} was substantially less favorable for the protein-compound complexes than those of the native protein-protein complex considering the much larger surfaces of the latter. However, ΔE_{SURF} for protein-protein and protein-compound complexes were similar for XIAP \cdot Smac and some of the Bcl-xL \cdot Bak antagonists (compounds **4** and **6**). This is explained by the fact that the XIAP \cdot Smac interface is relatively small, such that the protein-protein and protein-compound interfaces are similar in size. For the Bcl-xL \cdot Bak interaction inhibitors, compounds **4** and **6** had the most favorable ΔE_{VDW} (-83.3 ± 0.2 and -84.2 ± 0.1 kcal \cdot mol $^{-1}$, respectively) and ΔE_{SURF} (-11.1 and -9.8 kcal \cdot mol $^{-1}$, respectively).

The electrostatic contributions to the free energy of binding are represented by the Coulomb potential energy (ΔE_{ELE}) and the Generalized-Born (GB) solvation energy (ΔE_{GB}). The Coulomb energy is generally most favorable for the native protein ligands when compared to compounds. The only exception was for Bcl-xL \cdot Bak where two compounds, **2** and **3**, exhibited substantially more favorable ΔE_{ELE} . Compounds **4** and **6** also had highly favorable ΔE_{ELE} . This is likely due to the formation of a salt bridge between carboxylic groups on **4** and **6** with Arg-132 and Arg-139 on Bcl-xL, respectively. The favorable ΔE_{ELE} values lead to highly unfavorable ΔE_{GB} , since the desolvation of charged and polar groups is highly unfavorable. The Coulomb energies for the BRD4 \cdot H4 compounds were the most unfavorable. This may be attributed to the fact that none of these compounds have charged groups.

The ΔE_{GBTOT} term is the sum of polar and non-polar interactions. When entropy is added to ΔE_{GBTOT} , the result is the $\Delta G_{MM-GBSA}$ free energy of binding. As expected, $\Delta G_{MM-GBSA}$ is substantially less favorable than ΔE_{GBTOT} since the entropy for binding always results in a penalty to the free energy of binding. In most of the complexes, the entropy change due to binding of the native protein to the receptor was about 30 kcal \cdot mol $^{-1}$. For compounds, the entropy penalty was

more substantial for the Bcl-xL antagonists, which in some cases were nearly as large as that of the native ligand (e.g. **4**). This is because the Bcl-xL compounds are generally larger than the other compounds, but also possess linear architecture that makes them more flexible with more rotatable bonds. The larger number of rotatable bonds will result in more unfavorable entropy change following binding.

The computational ($\Delta G_{\text{MM-GBSA}}$) and experimental free energies (ΔG_{Exp}) of both protein-protein and protein-compound complexes are also shown in **Fig. 2.1**. The correlation coefficients when considering both protein-protein and protein-compound complexes are $r = 0.55$, $\rho = 0.43$, $\tau = 0.31$. When we only consider protein-compound complexes, the correlation coefficients are higher ($r = 0.64$, $\rho = 0.52$, $\tau = 0.38$). Among the individual components of the computational free energy, the total enthalpy (ΔE_{GBTOT}) components of the computational free energy correlates with the experimental free energy for all complexes. The antagonists of the Bcl-xL•Bak and BRD4•H4 interactions show the strongest correlations, with Pearson's r of 0.86 and 0.74, respectively. When entropy is considered ($\Delta G_{\text{MM-GBSA}}$), the correlation of Bcl-xL•Bak and MDM2•p53 remain relatively similar, while the correlation of XIAP•Smac and BRD4•H4 are lower by approximately 0.2. The predicted and experimental binding affinities did not correlate for the IL-2•IL-2R α complex.

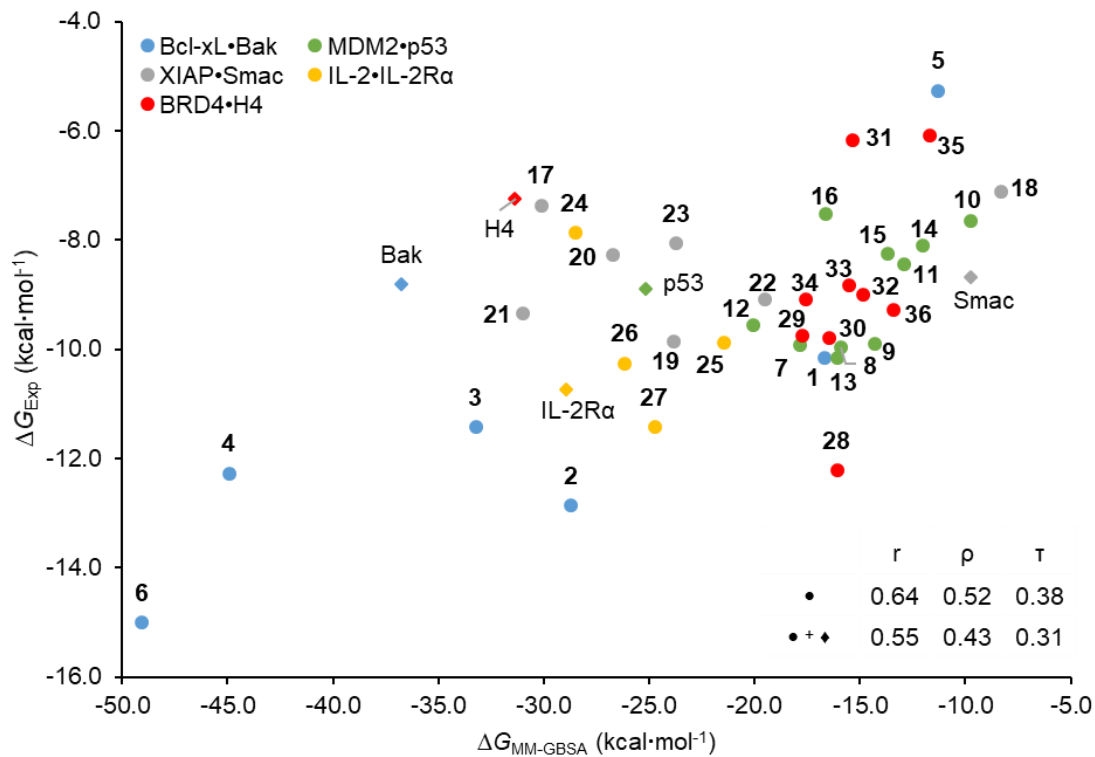


Figure 2.1. Comparison of free energies in protein-protein and protein-compound complexes. Free energies of protein-protein and protein-compound complexes. Protein-protein complexes are shown as diamonds while protein-compound complexes are shown in circles.

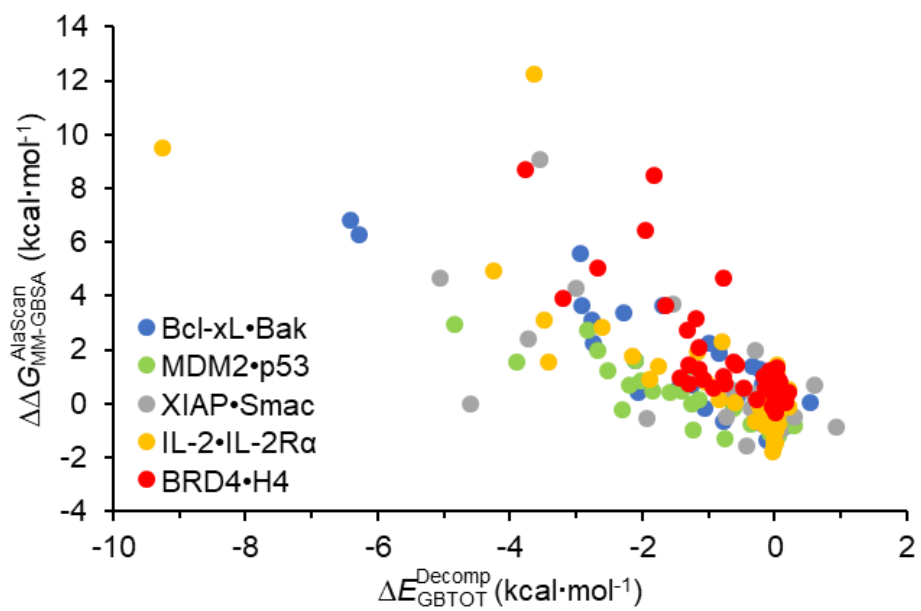


Figure 2.2. Per-residue decomposition versus computational alanine scanning of protein-protein complexes.

2.2.3 Computational Alanine Scanning and Free Energy Decomposition. To explore how effectively the native protein ligand engage receptor hot spots, we performed computational alanine scanning of interface residues on the receptor of each of the five protein-protein complexes using MM-GBSA. We mutated each receptor residue that is located at the interaction interface to alanine and calculated the resulting change in the MM-GBSA free energy between the wild-type and mutant complexes ($\Delta\Delta G_{\text{MM-GBSA}}^{\text{AlaScan}}$). In addition to alanine scanning, we carried out per-residue decomposition energy analysis for the native ligand and small-molecule inhibitors. This consists of determining the interaction energy of protein ligands and small-molecule inhibitors to each of the residues on the receptor. The decomposition energy includes all the components of the MM-GBSA free energy except for entropy.

We compare decomposition energies ($\Delta E_{\text{GBTOT}}^{\text{Decomp}}$) to alanine scanning free energy changes ($\Delta\Delta G_{\text{MM-GBSA}}^{\text{AlaScan}}$) for each residue at the interaction interface (**Fig. 2.2**). We observe good correlation between the computational alanine scanning and total residue decomposition energies across the five protein-protein complexes. The mean correlation coefficient across the five complexes for r , ρ , and τ are -0.80 ± 0.04 , -0.53 ± 0.06 , and -0.38 ± 0.05 , respectively. A negative correlation is expected since residues that have favorable decomposition energies (negative energies) with the protein ligand are expected to lead to a higher binding affinity penalty (positive energy) when mutated to alanine. Inspection of **Fig. 2.2** reveals that there are several exceptions. Several residues lead to unfavorable $\Delta\Delta G_{\text{MM-GBSA}}^{\text{AlaScan}}$, yet their interaction with the protein ligand ($\Delta E_{\text{GBTOT}}^{\text{Decomp}}$) is highly favorable.

2.2.4 Bcl-xL•Bak. The Bcl-2 protein family consists of apoptosis regulators, which are divided into three subfamilies: pro-survival (e.g. Bcl-xL, Bcl-w, and Mcl-1), Bax-like pro-apoptotic (e.g. Bax and Bak), and BH3 only (e.g. Bad and Bim) [95]. Inhibition of pro-survival activity occurs through binding of a BH3 domain to a hydrophobic cleft formed by the BH1, BH2, and BH3 domains of a pro-survival protein [96, 97]. One example is the complex between the pro-survival protein Bcl-xL and the pro-apoptotic protein Bak, which is characterized by a 16-residue α -helix peptide in a hydrophobic cleft formed by the BH1, BH2, and BH3 regions of Bcl-xL (**Fig. 2.3A**). This 16-residue peptide, with a binding affinity of $0.34 \mu\text{M}$, represents the minimal region required to bind to Bcl-xL [57].

Experimental alanine scanning of the Bak peptide identified Val-74, Arg-76, Leu-78, Ile-81, Asp-83, and Ile-85 on Bak, and Arg-139 on Bcl-xL as critical for the interaction, while Ile-80 and Asp-84 on Bak and Arg-100 on Bcl-xL were not as important [57]. No experimental alanine scanning was done on the receptor. Thus, we carried out a computational alanine scan using MM-

GBSA to identify residues on Bcl-xL that are critical for binding to Bak (**Fig. 2.3B**). We found that the computational mutation of nine residues to alanine resulted in more than 1.5 kcal·mol⁻¹ increase in the MM-GBSA free energy. These predicted hot spots include Phe-97, Arg-100, Tyr-101, Phe-105, Leu-108, Glu-129, Leu-130, Arg-139, Phe-146, and Tyr-195. Arg-139 was predicted to be critical for binding consistent with experimental data [57].

We carried out decomposition energy calculations to determine the interaction energy between Bak and individual residues on Bcl-xL (**Fig. 2.3C**). There was strong engagement of hot spots by Bak as evidenced by $\Delta E_{\text{GBTOT}}^{\text{Decomp}}$ magnitudes that were overall greater than 2 kcal·mol⁻¹. The only exceptions are Phe-146 and Tyr-195, which interact with Bak with decomposition energies of -0.83 ± 0.01 and -1.69 ± 0.04 kcal·mol⁻¹, respectively. We also found that Bak engaged some residues that are not considered hot spots. For example, Val-126 binds to Bak with a decomposition energy of -2.06 ± 0.02 kcal·mol⁻¹ despite mutation of Val-126 that resulted in a mere 0.42 kcal·mol⁻¹ change in MM-GBSA energy. Interestingly, this hydrophobic residue is among residues on Bcl-xL that form contacts with BH3-containing antagonist peptides [97]. Work by Oberstein and co-workers detailed the differences in van der Waals contacts of two BH3 peptides, Beclin-1 and Bim, at Tyr-101 and Leu-108 of Bcl-xL [98]. They suggest that these differences were critical for the binding specificity of BH3 peptides to Bcl-xL [98]. Previous mutagenesis studies of Val-126 against other BH3 peptides revealed the importance of the residue in heterodimerization of Bcl-xL [99, 100].

Decomposition energies for small molecules (**1** to **6**) were carried out to gain insight into their engagement of individual hot spots on Bcl-xL (**Fig. 2.4A**). We compared decomposition energies of small molecules with those of the native Bak peptide to uncover how effectively compounds mimic the native ligand. Surprisingly, in most cases, compounds do not engage Bcl-xL hot spots as effectively as Bak, despite the substantial medicinal chemistry efforts that were invested in developing these compounds. For example, none of the compounds show $\Delta E_{\text{GBTOT}}^{\text{Decomp}}$ values that are equal or greater than those of Bak for hot spot residues Arg-100, Tyr-101, and Glu-129. The remaining hot spots, Phe-97, Phe-105, Leu-130, Arg-139, and Tyr-195 engage four, one, five, two, and three of the six compounds with similar interaction energies to the native protein Bak, respectively. We found three hot spots bind strongly to all three of the nanomolar inhibitors, namely Phe-97, Leu-130, and Tyr-195. The sub-nanomolar inhibitor **6** shows a unique pattern of hot-spot binding. The compound binds to Phe-105, Leu-108 and Arg-139 much more strongly than the other compounds (-3.06 ± 0.01 , -3.51 ± 0.02 , and -8.15 ± 0.04 kcal·mol⁻¹, respectively). In each of these cases, compound **6** binds to these hot spots much more strongly than that of the native ligand Bak.

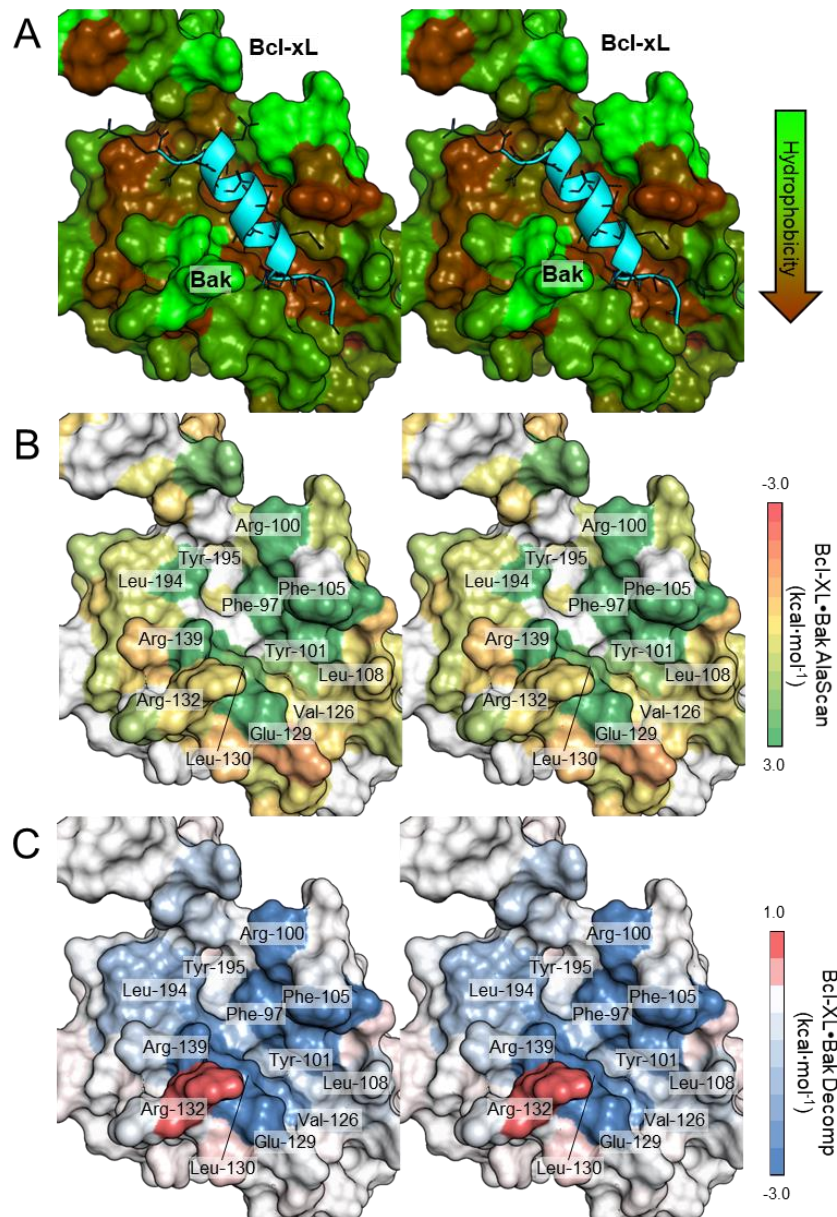


Figure 2.3. Bcl-xL•Bak Protein-Protein Complex. (A) The protein complex of Bcl-xL and Bak peptide. Bcl-xL is shown in surface and colored by hydrophobicity, with more hydrophobic residues in brown and more hydrophilic residues in green. The Bak peptide is shown in cyan and represented in cartoon with side chains in stick. (B) Surface representation of Bcl-xL, where residues at the interface on Bcl-xL are colored based on the change in free energy after mutating the residue to alanine. (C) Surface representation of Bcl-xL colored by per-residue decomposition energy with Bak.

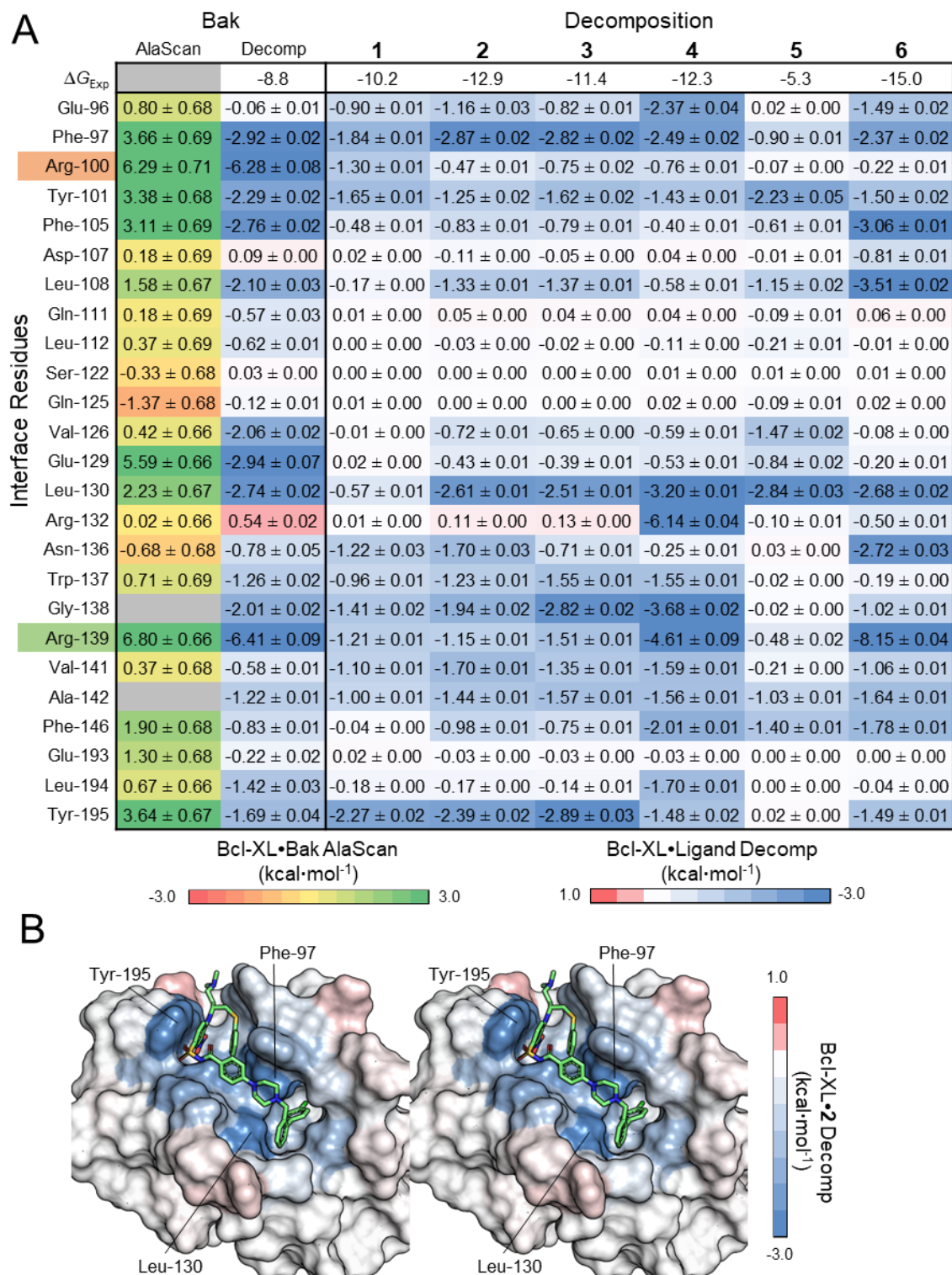


Figure 2.4. Bcl-xL•Bak comparison with inhibitors. (A) Residues on Bcl-xL at the interface of the protein-protein interaction. On the left, residues are color-coded based on experimental mutagenesis studies. Known hot spots residues are highlighted in green and residues that are not

hot spots are highlighted in orange. The first two columns show alanine scanning and per-residue decomposition at the specific residue in the protein-protein complex, respectively. The third to last column show the per-residue decomposition of at the specific residue for each co-crystallized inhibitor. Alanine scanning and per-residue decomposition energies are color-coded per the scales in **Fig. 2.3B** and **Fig. 2.3C**, respectively. Experimental ΔG are shown in the first row for each complex. Numbers are shown in kcal·mol⁻¹. **(B)** Surface representation of Bcl-xL colored by per-residue decomposition energy with compound **2**.

Individual compounds showed differences in their binding to hot spots when compared to the native peptide Bak. Compound **5**, which has poor micromolar affinity, binds weaker to hot spots compared to Bak, particularly at Phe-97, Arg-100, Phe-195, and Asn-136 to Arg-139. Compound **1**, on the other hand, interacts more tightly with Arg-100 than the other compounds, but shows little interaction with hot spots Leu-108, Val-126, Glu-129, Leu-130, and Phe-146. The co-crystallized structure of **2** is shown in **Fig. 2.4B**. Compound **2** (ABT-737) binds to Bcl-xL, Bcl-2, and Bcl-w with sub-nanomolar affinities, but shows micromolar affinities to Mcl1 [58]. Critical interactions between Bcl-xL and **2** include the π - π and π -cation interactions between Tyr-195 and nitrobenzene of **2**, as well as hydrogen bonding between Gly-138 and Asn-136 to the sulfone and nearby secondary amine moieties of **2**, respectively. Tyr-101 also forms a π - π interaction with another benzene ring in the core structure of the compound. The decomposition energy between Tyr-101 and **5** is similar to that of the native peptide and is an additional 1 kcal·mol⁻¹ better than the other compounds. Finally, Phe-97 and Val-141 form hydrophobic contacts with the thiophenol of **2**. The interaction with Phe-97 in **1** and **5** is much weaker than in the other compounds, despite π - π interactions with the residue in both compounds. Compounds **3** and **4** share similar core structures and binding modes with **2**. A modification an amide carbonyl in the core of **2** to generate a quinazoline in **3** results in an additional hydrogen bond between the quinazoline and the side chain of Tyr-101. Compound **4** forms a salt bridge with Arg-139. Among all six antagonists of Bcl-xL, only **4** and the sub-nanomolar compound **6** interact with the key Arg-139 residue. The most potent compound, **6**, forms a π - π interaction with Phe-105 and hydrogen bonds with Leu-108 and Asn-136. These additional interactions allow the compound to engage more hot spots on Bcl-xL and mimic more of the interactions seen in the native Bak peptide.

2.2.5 MDM2•p53. MDM2 is an inhibitor of transcriptional activity of the tumor suppressor p53 [101]. This interaction is characterized by a 15-residue α -helix of p53 binding into a hydrophobic cleft of MDM2 (**Fig. 2.5A**). The region on p53 from Thr-18 to Leu-26 represents the minimal region required to bind to MDM2 [102]. On p53, the three side chains of Phe-19, Trp-23, and Leu-26 are buried in the hydrophobic pocket of MDM2 and are critical for binding [103]. There are no published alanine scanning studies for MDM2 *in vitro*, therefore, the contributions of residues at the MDM2•p53 interface to binding of the p53 peptide are unknown. This prompted us to conduct a computational alanine scan to identify hot spots at the interface (**Fig. 2.5B**). We find that Met-50, Leu-54, Ile-61, Val-93, and Tyr-100 on MDM2 are hot spots, resulting to more than 1.5 kcal·mol⁻¹ change in $\Delta\Delta G_{MM-GBSA}^{AlaScan}$. Similarly, the decomposition energies show that Thr-26, Met-50, Leu-51, Leu-54, Ile-61, Val-93, Arg-97, Tyr-100, and Tyr-104 contribute more than 2 kcal·mol⁻¹ to the interaction (**Fig. 2.5C**).

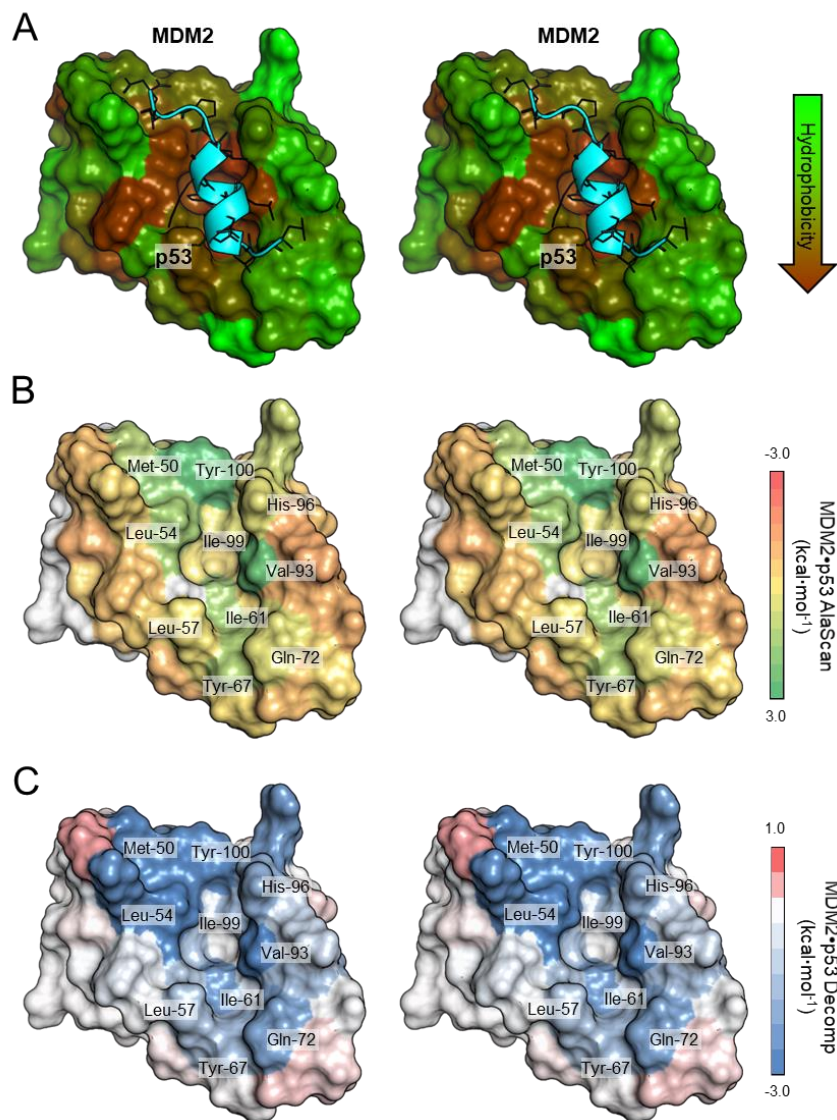


Figure 2.5. MDM2•p53. **(A)** The protein complex of MDM2 and p53 peptide. MDM2 is shown in surface and colored by hydrophobicity, with more hydrophobic residues in brown and more hydrophilic residues in green. The p53 peptide is shown in cyan and represented in cartoon with side chains in stick. **(B)** Surface representation of MDM2, where residues at the interface on MDM2 are colored based on the change in free energy after mutating the residue to alanine. **(C)** Surface representation of MDM2 colored by per-residue decomposition energy with p53.

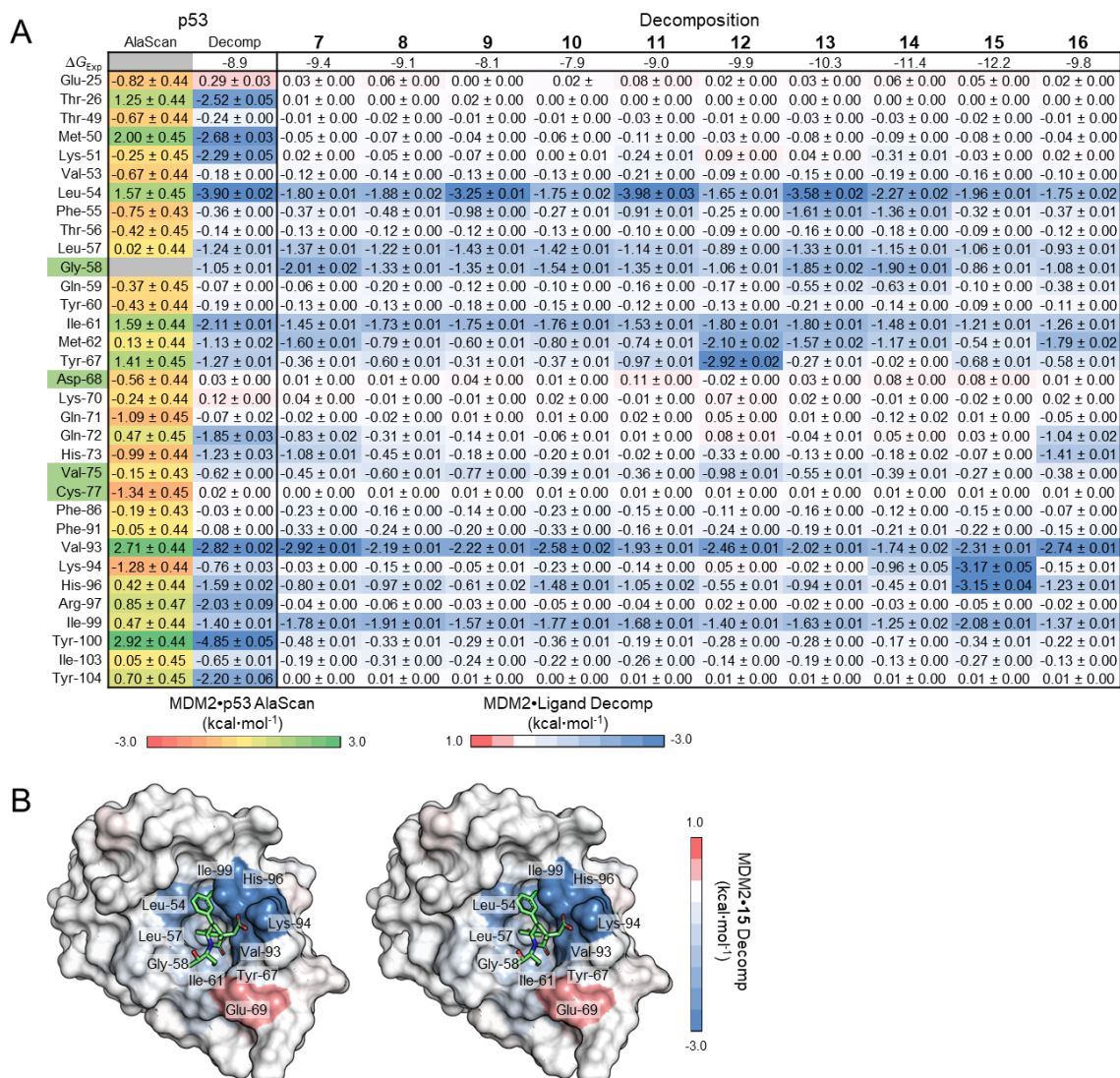


Figure 2.6. MDM2•p53 comparison with inhibitors. **(A)** Residues on MDM2 at the interface of the protein-protein interaction. On the left, residues are color-coded based on experimental mutagenesis studies. Known hot spots residues are highlighted in green. The first two columns show alanine scanning and per-residue decomposition at the specific residue in the protein-protein complex, respectively. The third to last column show the per-residue decomposition of at the specific residue for each co-crystallized inhibitor. Alanine scanning and per-residue decomposition energies are color-coded per the scales in **Fig. 2.5B** and **Fig. 2.5C**, respectively. Experimental ΔG are shown in the first row for each complex. Numbers are shown in kcal·mol⁻¹. **(B)** Surface representation of MDM2 colored by per-residue decomposition energy with compound **15**.

Decomposition energies for small molecules (**7** to **16**) were carried out to gain insight into their binding to hot spots on MDM2 (**Fig. 2.6A**). Three hot spots that were found to be strongly engaged by p53 bind strongly to most compounds, namely Leu-54, Ile-61, and Val-93. It is worth noting, however, that Ile-61 interacts with compounds worse than the native p53 protein. Two of the hot spots that are strongly engaged by p53 did not bind to compounds **7-16**. For example, p53 binds to Met-50 and Tyr-100 with decomposition energies of -2.68 ± 0.03 and -4.85 ± 0.05 kcal·mol⁻¹, respectively. Yet, the magnitude of the interaction energies of the compounds to these hot spots is nearly consistently below 1 kcal·mol⁻¹. It is interesting to note that p53 strongly binds to Tyr-100 with a decomposition energy of -4.85 ± 0.05 kcal·mol⁻¹. Some compounds, like the most potent inhibitor, namely **15**, revealed unique interactions not present in others. The compound binds very strongly to Lys-94 and His-96 with decomposition energies absolute values greater than 3 kcal·mol⁻¹. Like the p53 peptide, compounds **7-16** show strong interactions to Leu-54 and Val-93, and weaker binding to Leu-57, Gly-58, Ile-61, Met-62, His-96, and Ile-99. The majority fail to mimic the interaction energies of p53 at Tyr-67, Gln-72, His-73, Arg-97, Tyr-100, and Tyr-104.

Compound **15** (AM-8553), shown in **Fig. 2.6B**, binds to MDM2 with an affinity of 0.4 nM [72]. One of the chlorobenzenes of **15** occupies p53's Leu-26 pocket and forms a π - π interaction with the imidazole of His-96. The carboxylic acid of **15** forms a salt bridge with Lys-94 and a hydrogen bond with His-96. The interactions between the compound and Lys-94 and His-96 on MDM2 is reflected in the favorable -3 kcal·mol⁻¹ interaction decomposition energies. These two interactions are absent in the other (weaker) compounds. The other chlorobenzene group mimics the six-membered ring of the indole of p53's Trp-23. The hydroxyl group points away from the positively charged Glu-69, thereby allowing the nearby ethyl group to occupy p53's Phe-19 pocket and engage MDM2's Gly-58, Ile-61, and Met-62. Replacement of the hydroxyl group of **15** with a sulfone resulted in a 10-fold improvement in K_d [104]. By mimicking both Trp-23 and Leu-26 on p53, compounds engage favorably with Leu-54 and Ile-99. Similarly, mimicking both Phe-19 and Trp-23 on p53 engages Ile-61 and Val-93 on the receptor. In sum, it appears that the MDM2 antagonists have been designed to mimic hot spots of p53.

2.2.6 XIAP•Smac. The E3 ubiquitin-protein ligase XIAP is a member of the Inhibitor of Apoptosis Proteins (IAP) family, which suppresses apoptotic cell death pathways through the inhibition of caspases [105, 106]. Smac/DIABLO binding at the protein-protein interface on the BIR3 domain of XIAP interferes with XIAP inhibition of CASP9 [74, 107]. A short AVPI-peptide at the N-terminal of Smac forms the interface at the dimer structure between the protein and the BIR3 domain of XIAP (**Fig. 2.7A**). Mutation of any of the first four residues of the Smac peptide resulted in greater than 100-fold decrease in binding affinity of the BIR3 of XIAP [74, 108]. On

XIAP, mutations at Asp-296, Leu-307, Trp-310, Glu-314, and Trp-323 greatly decreased the binding affinity of Smac, while mutations at Asp-315, Glu-318, His-343, and Gln-319 had little to no effect [74]. Although mutation of His-343 had little effect on Smac binding, *in vitro* inhibition of caspase-9 was completely abrogated [74].

On XIAP, mutations to alanine at Arg-258, Leu-307, Trp-310, Glu-314, Trp-323, and Tyr-324 resulted in more than $1.5 \text{ kcal}\cdot\text{mol}^{-1}$ change in free energy and are considered hot spots (**Fig. 2.7B**). We also carried out decomposition analysis to explore whether Smac binds strongly to hot spots. We found that Arg-258, Leu-307, Thr-308, Glu-314, and Trp-323 contributed more than $2 \text{ kcal}\cdot\text{mol}^{-1}$ to the interaction energy with Smac while Gly-306, Asp-309, and Trp-310 bound more weakly to Smac with approximately $1 \text{ kcal}\cdot\text{mol}^{-1}$ in the energy decomposition (**Fig. 2.7C**). In the XIAP•Smac dimer, Arg-258 forms a salt bridge with Glu-9 on Smac. However, iterative truncation of the first nine residues of the Smac peptide down to the first five residues did not affect the binding affinity to the BIR3 domain of XIAP [74]. Therefore, it is unlikely that Arg-258 is critical to XIAP•Smac binding and inhibition. Gly-306 and Thr-308 are native lysine residues in the BIR2 domain of XIAP, and may account for the differences in binding affinity between the BIR2 and BIR3 domains [74]. While Gly-306 cannot be tested through alanine scanning, computational mutation of Thr-308 to alanine resulted in negligible change in the binding free energy and are not considered hot spots.

Decomposition energies were determined for **17-23** to compare with interaction energies of the Smac native ligand (**Fig. 2.8A**). Among the six hot spots that we found on XIAP (Arg-258, Leu-307, Trp-310, Glu-314, Trp-323, and Tyr-324), three interact strongly with the compounds, namely Leu-307, Glu-314 and Trp-323. These three amino acids show the tightest binding to both Smac and compounds. Trp-310 binds to Smac with a decomposition energy of $-1.53 \pm 0.01 \text{ kcal}\cdot\text{mol}^{-1}$, which is relatively weak. However, most compounds interact with this residue as strongly as Smac, except for **18**, which interestingly is one of the weaker compounds with an experimental binding affinity of $-6.2 \text{ kcal}\cdot\text{mol}^{-1}$. Arg-258 and Tyr-324 are hot spots that do not bind to any of the compounds, even though Arg-258 shows very strong interaction to Smac. Further inspection of the data reveals Thr-308 binds strongly to all compounds as evidenced by decomposition energies that are on average $-5 \text{ kcal}\cdot\text{mol}^{-1}$. Asp-309 was not as critical to the binding since two of the most potent inhibitors, namely **17** and **23**, do not engage this residue with high affinity.

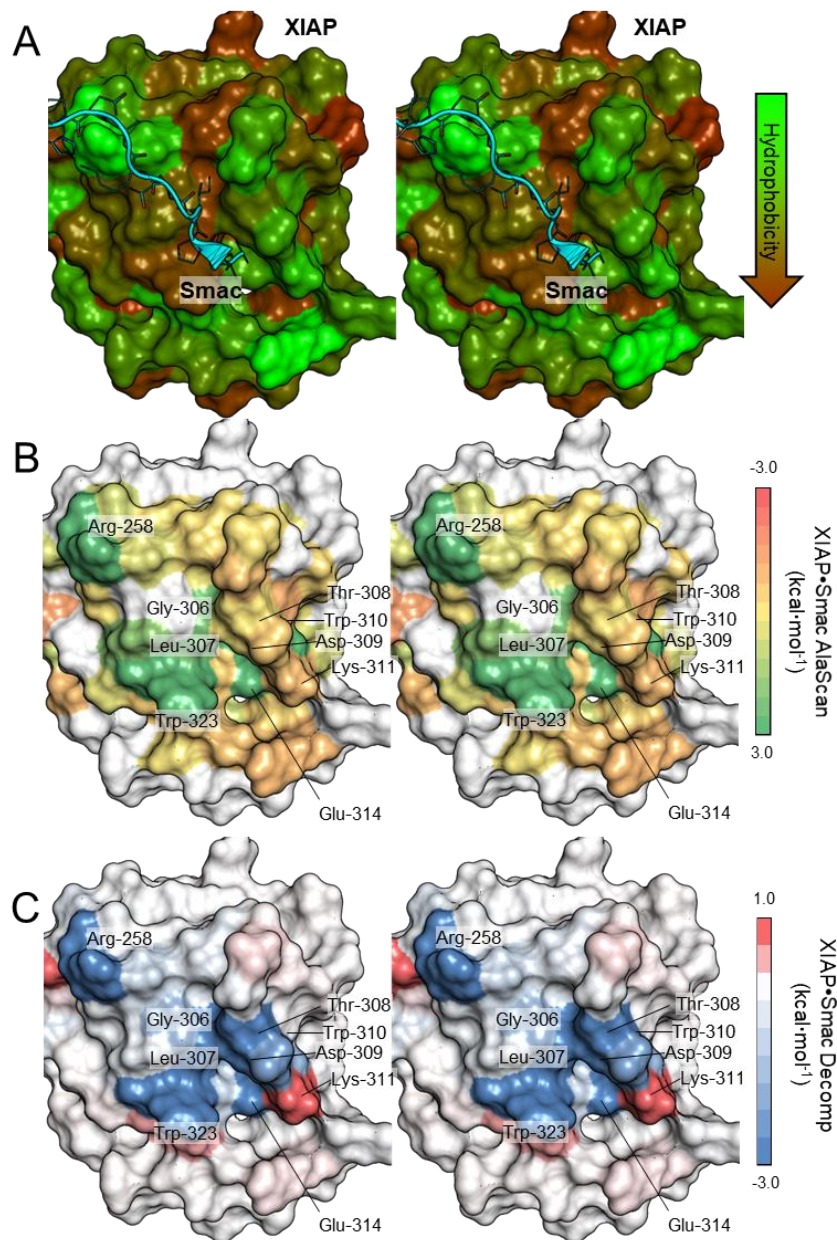


Figure 2.7. XIAP•Smac. **(A)** The protein complex of XIAP and Smac/DIABLO. XIAP is shown in surface and colored by hydrophobicity, with more hydrophobic residues in brown and more hydrophilic residues in green. Smac is shown in cyan and represented in cartoon with side chains in stick. **(B)** Surface representation of XIAP, where residues at the interface on XIAP are colored based on the change in free energy after mutating the residue to alanine. **(C)** Surface representation of XIAP colored by per-residue decomposition energy with Smac.

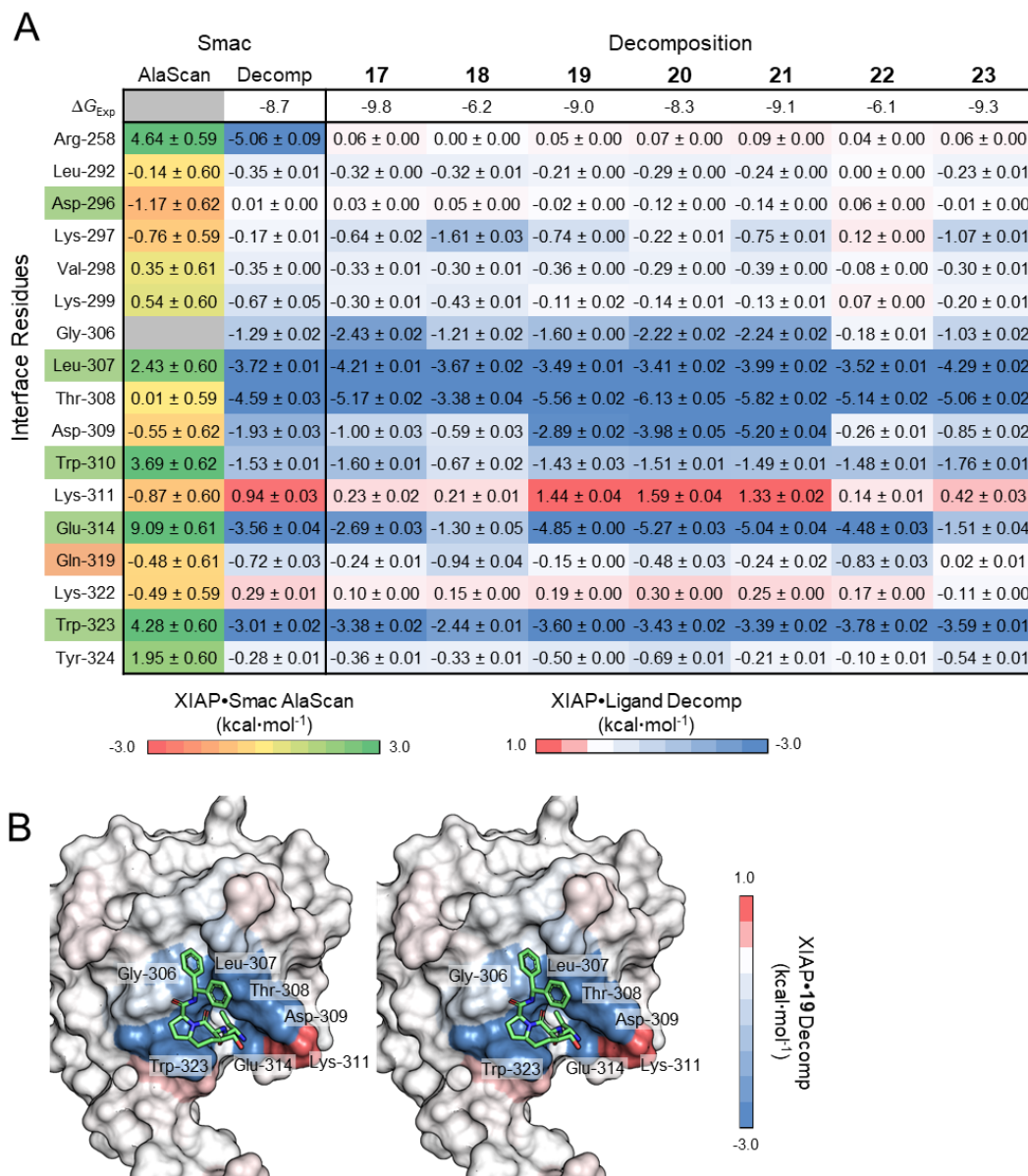


Figure 2.8. XIAP•Smac comparison with inhibitors. **(A)** Residues on XIAP at the interface of the protein-protein interaction. On the left, residues are color-coded based on experimental mutagenesis studies. Known hot spots residues are highlighted in green and residues that are not hot spots are highlighted in orange. The first two columns show alanine scanning and per-residue decomposition at the specific residue in the protein-protein complex, respectively. The third to last column show the per-residue decomposition of at the specific residue for each co-crystallized inhibitor. Alanine scanning and per-residue decomposition energies are color-coded per the scales in **Fig. 2.7B** and **Fig. 2.7C**, respectively. Experimental ΔG are shown in the first row for each complex. Numbers are shown in kcal·mol⁻¹. **(B)** Surface representation of XIAP colored by per-residue decomposition energy with compound **19**.

In the protein-protein complex, the bulk of the contribution at Glu-314 is from electrostatic interactions with the N-terminal Ala-1 residue of the Smac peptide. Among the compounds, there is a common amine moiety that replicates this interaction. The binding mode of **19** is shown in **Fig. 2.8B**. Compounds **19-21** differ only by a hydroxymethyl, methylamine, and ethylamine substituent on the seven-membered ring, respectively. This results in an approximately $1.1 \text{ kcal}\cdot\text{mol}^{-1}$ stronger interaction with Asp-309 on XIAP. While, the residue is not critical to Smac binding, and the K_d and IC_{50} of **20** is about four-fold weaker than the other two compounds. The fused cycloheptane and pyrrolidine ring mimic the side chains of Val-2 and Pro-3 of the Smac peptide, respectively. The fused ring also serves to bury the hydrophobic Leu-307 and Trp-323 residues. The final critical residue of the Smac peptide, Ile-4, is mimicked by one of the two benzene rings of the compound. While not hot spots, the compound is stabilized by hydrogen bonds between the backbone of Gly-306 with the acetylamide by the two benzene rings of the compound and Thr-308 with the aminobutanamide substituent of the fused ring. The two micromolar compounds, **18** and **21**, fail to mimic Val-2 and Ile-4 of the Smac peptide, respectively.

2.2.7 IL-2•IL-2R α . IL-2 is produced after antigen activation during an immune response and binds to a combination of α -, β -, and γ - IL-2 receptors [109]. While each receptor subunit can bind to IL-2 at varying affinities, ranging from approximately 10 nM for the α -subunit to about 0.7 mM for the γ subunit, the tetramer complex is approximately 5 pM [48]. Here, we explore the heterodimeric interface between IL-2 and its α -subunit (**Fig. 2.9A**). Site directed mutagenesis and other hot spot identification techniques have identified Lys-35, Arg-38, Phe-42, Lys-43, Tyr-45, Glu-62, and Leu-72 as critical residues on IL-2 at the interface [110-112]. Comparative mutagenesis of IL-2 with both the α -subunit and compound **27** showed that Phe-42, Tyr-45, and Glu-62 were critical for binding, while mutations Met-39, Thr-41, Lys-43, Phe-44, and Leu-72 showed moderate disruption in binding affinity of subunit binding [84]. However, mutations at Lys-35 and Arg-38 showed less than 5-fold change in binding affinity and mutations at Pro-65 and Val-69 were negligible [84].

Both the alanine scanning (**Fig. 2.9B**) and residue decomposition (**Fig. 2.9C**) analyses of the IL-2•IL-2R α largely replicate the experimental mutagenesis of interface residues. Residues that contribute more than $2 \text{ kcal}\cdot\text{mol}^{-1}$ to the binding affinity in the residue decomposition include Lys-35, Arg-38, Phe-42, Lys-43, Tyr-45, Glu-62, and Pro-65. Along with these residues, Phe-44 and Glu-61 are hot-spots residues from the alanine scanning that resulted in greater than $1.5 \text{ kcal}\cdot\text{mol}^{-1}$ difference in binding affinity when computationally mutated. Despite the nearly 3-fold decrease in binding affinity from experimentally mutating Arg-38, the residue contributes about $-9 \text{ kcal}\cdot\text{mol}^{-1}$ to the decomposition and approximately $9.5 \text{ kcal}\cdot\text{mol}^{-1}$ change in free energy upon computational

mutation to alanine. In the complex, Arg-38, along with Lys-35 and Lys-43 are among a set of positively charged residues that interacts with a set of negatively charged residues on the α -subunit.

Decomposition energies for compounds **24-27** were determined to compare their interaction to IL-2•IL-2R α . Among the nine hot spots that were identified from our alanine scan, six of them engage compounds very strongly, namely Lys-35, Arg-38, Phe-42, Lys-43, Glu-62, and Leu-72. However, only four of these hot spots bind strongly to most compounds, namely Arg-38, Phe-42, Lys-43, and Glu-62. Lys-35 only binds to **26** with decomposition energy that is less than 2 kcal·mol⁻¹ in magnitude, although this is still weaker than the residue's binding to IL-2R α . It is interesting to note that compound **26** is the most potent inhibitor. Arg-38 shows substantial binding to IL-2R α and engages compounds with favorable decomposition energies around -2 kcal·mol⁻¹. These interactions remain substantially lower than the very strong interaction between Arg-38 and IL-2R α . One residue, namely Leu-72, shows relatively high affinity to most compounds, yet the residue was not found to be a hot spot in our alanine scan. Two hot spots, Tyr-45 and Phe-44 show weak interaction with the compounds. In the case of Phe-44, even IL-2R α binds weakly to the residue.

Compounds **24-27** interact with major hot spots on IL-2 in a similar manner to IL-2R α (**Fig. 2.10A**). Compared to the native subunit, the antagonists show similar interactions at Arg-38, Phe-42, Lys-43, Glu-62, and Leu-72, and weaker interaction energies at Lys-35, Tyr-45, and Pro-65. A common imine group mimics Arg-36 on the subunit and forms hydrogen bonds with the side chain of Glu-62. Another common carbonyl forms hydrogen bonds with the Lys-43 hot spot to stabilize the compound. At the opposite end of the compound, aromatic rings form salt bridges with Arg-38. The orientation on the hot spot Phe-42 side chain points down into the binding pocket when bound to the antagonists but points out when bound to the native α -subunit. The conformational change at this residue flattens the interface, allowing the compounds to adopt their respective binding modes. Compounds **25-27** are analogs with a common binding mode (**26** is shown in **Fig. 2.10B**). Compounds **26** and **27** feature additional substituents compared to their analogs **23** and **25**. Extending the core structures and adding carboxylic acid and amide moieties in **27** and **28**, respectively, allow the compounds to mimic Asp-4 on the native ligand and interact with the Lys-35 hot spot on IL-2. A π -cation interaction between Tyr-45 on IL-2 and Arg-35 on the native ligand is not seen in any of the compounds. A cyclohexane in **25** and isobutyl groups in **26** and **27** form weak contacts with Tyr-45, but the decomposition energy is less favorable than -1.3 kcal·mol⁻¹.

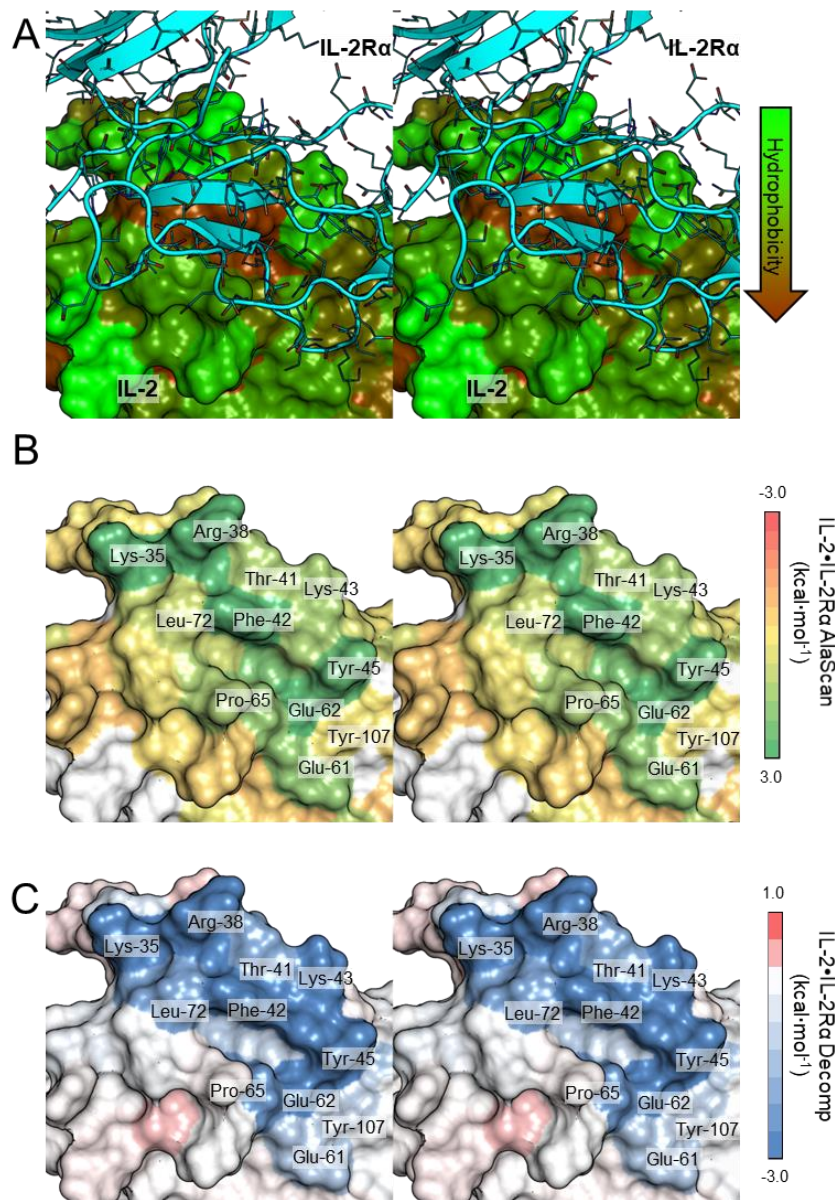


Figure 2.9. IL-2•IL-2R α . (A) The protein complex of IL-2 and IL-2R α . IL-2 is shown in surface and colored by hydrophobicity, with more hydrophobic residues in brown and more hydrophilic residues in green. IL-2R α is shown in cyan and represented in cartoon with side chains in stick. (B) Surface representation of IL-2, where residues at the interface on MDM2 are colored based on the change in free energy after mutating the residue to alanine. (C) Surface representation of IL-2 colored by per-residue decomposition energy with IL-2R α .

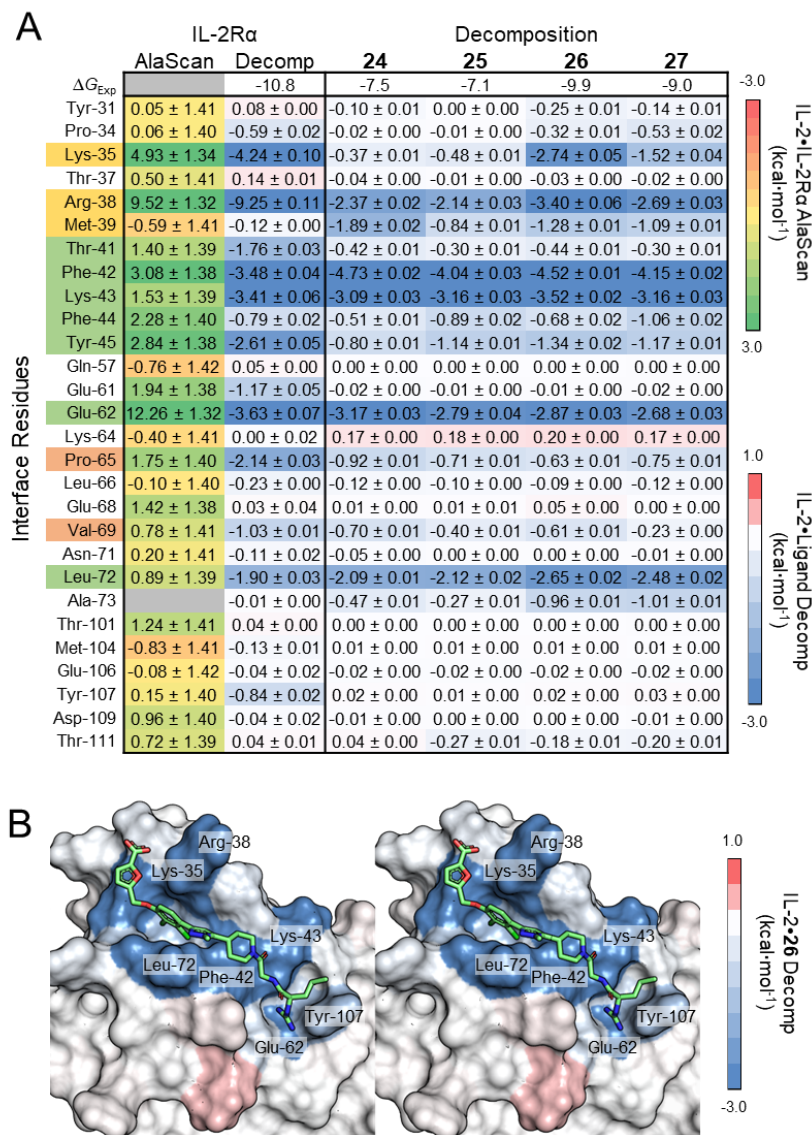


Figure 2.10. IL-2•IL-2R α comparison with inhibitors. **(A)** Residues on IL-2 at the interface of the protein-protein interaction. On the left, residues are color-coded based on experimental mutagenesis studies. Known hot spots residues are highlighted in green and residues that are not hot spots are highlighted in orange. Mutations that do not greatly impact the interaction are highlighted in yellow. The first two columns show alanine scanning and per-residue decomposition at the specific residue in the protein-protein complex, respectively. The third to last column show the per-residue decomposition of at the specific residue for each co-crystallized inhibitor. Alanine scanning and per-residue decomposition energies are color-coded per the scales in **Fig. 2.9B** and **Fig. 2.9C**, respectively. Experimental ΔG are shown in the first row for each complex. Numbers are shown in kcal·mol⁻¹. **(B)** Surface representation of IL-2 colored by per-residue decomposition energy with compound **26**.

2.2.8 BRD4•H4. The second primary interaction involves a prototypical member of the bromodomain family, BRD4, and an acetylated histone tail. The family contains 61 bromodomains on 46 proteins that affect post-translational modification by reading the acetylated lysines of epigenetic markers [113, 114]. The structure contains two acetylated lysine residues at Lys-5 and Lys-8 buried deep into a hydrophobic pocket on the first bromodomain of BRD4 (**Fig. 2.11A**). Alanine scanning of a tetra-acetylated H4 peptide revealed that residues immediately flanking the first two acetylated sites, Lys-5(ac) and Lys-8(ac), were critical (i.e. Gly-4, Gly-6, Gly-9, and Leu-10) [85]. On the interface between BRD4 and Lys-5(ac) and Lys-8(ac) H4, mutations at Trp-81, Leu-94, Tyr-97, Asn-140, Asp-145, and Met-149 resulted in approximately ten-fold reduction in binding activity, while mutations at Pro-82, Tyr-139, Asp-144, and Ile-146 resulted in approximately two-fold reduction in binding activity [85].

In the computational alanine scan, we identify several residues as potential hot spots (**Fig. 2.11B**). These include Phe-79, Val-87, Leu-94, Asp-96, Tyr-139, Asn-140, Lys-141, Asp-144, Asp-145, Ile-146, and Met-149. While mutation of Ile-146 only reduced binding activity of the double acetylated peptide by two-fold, the residue contacts both acetylated residues in the complex and contributes significantly to both histone and compound interactions. In the residue decomposition, we identify the trio of Tyr-139, Asn-140, and Ile-105 contributing more than 2 kcal·mol⁻¹ to the interaction between BRD4 and double acetylated H4 (**Fig. 2.11C**). Although Trp-81 was identified as a critical residue, it only contributes approximately 0.76 kcal·mol⁻¹ in both the residue decomposition and alanine scanning. In the crystal structure, the residue is solvent exposed and shields Lys-8(ac) from the solvent.

Decomposition energies for compounds **28-36** reveal that most hot spots on BRD4 do not effectively engage the bound small molecules. For example, Phe-79, Asp-96, Tyr-139, Asp-144, Asp-145, and Met-149 show no interaction with the compounds. In fact, Asp-96 and Tyr-139 appear to be critical for H4 binding to BRD4 as evidenced by the loss of more than 8 kcal·mol⁻¹ upon their mutation to alanine. Inspection of **Fig. 2.11A** shows that these residues are located outside the binding pocket that is occupied by the BRD4 antagonists. Compound substituent that bind outside the binding pocket occupy sites that contain Trp-81 and Ile-146, both of which do not contribute as much to H4 binding as Asp-96 and Tyr-139. Several amino acids that are not hot spots showed strong binding to compounds. Pro-82 binds to compounds **33** and **36** with decomposition energies that are more favorable than -2 kcal·mol⁻¹. Another example is Leu-92, which showed a penalty of 1.44 ± 0.49 kcal·mol⁻¹ upon mutation to alanine and engaged most compounds with decomposition energies that are less than -2 kcal·mol⁻¹.

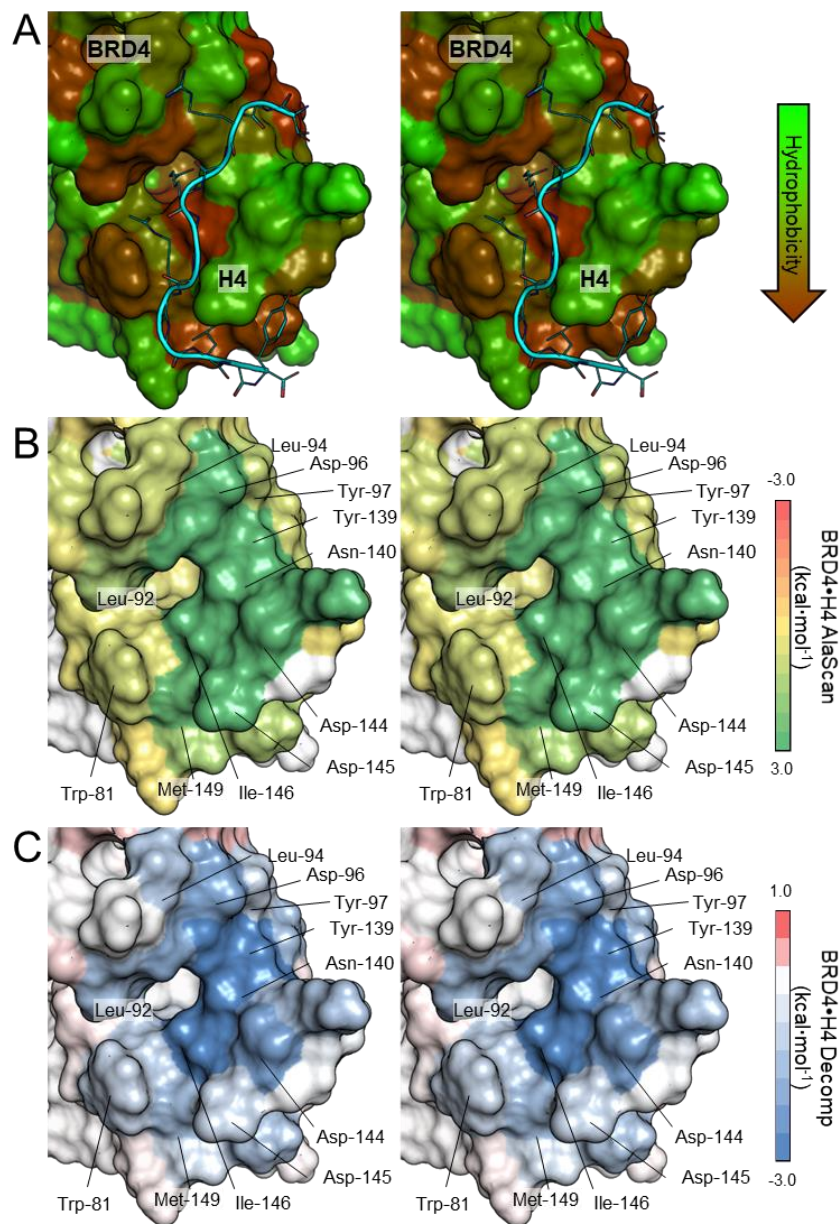


Figure 2.11. BRD4•H4. (A) The protein complex of BRD4 and H4 peptide. BRD4 is shown in surface and colored by hydrophobicity, with more hydrophobic residues in brown and more hydrophilic residues in green. The H4 peptide is shown in cyan and represented in cartoon with side chains in stick. (B) Surface representation of BRD4, where residues at the interface on BRD4 are colored based on the change in free energy after mutating the residue to alanine. (C) Surface representation of BRD4 colored by per-residue decomposition energy with H4.

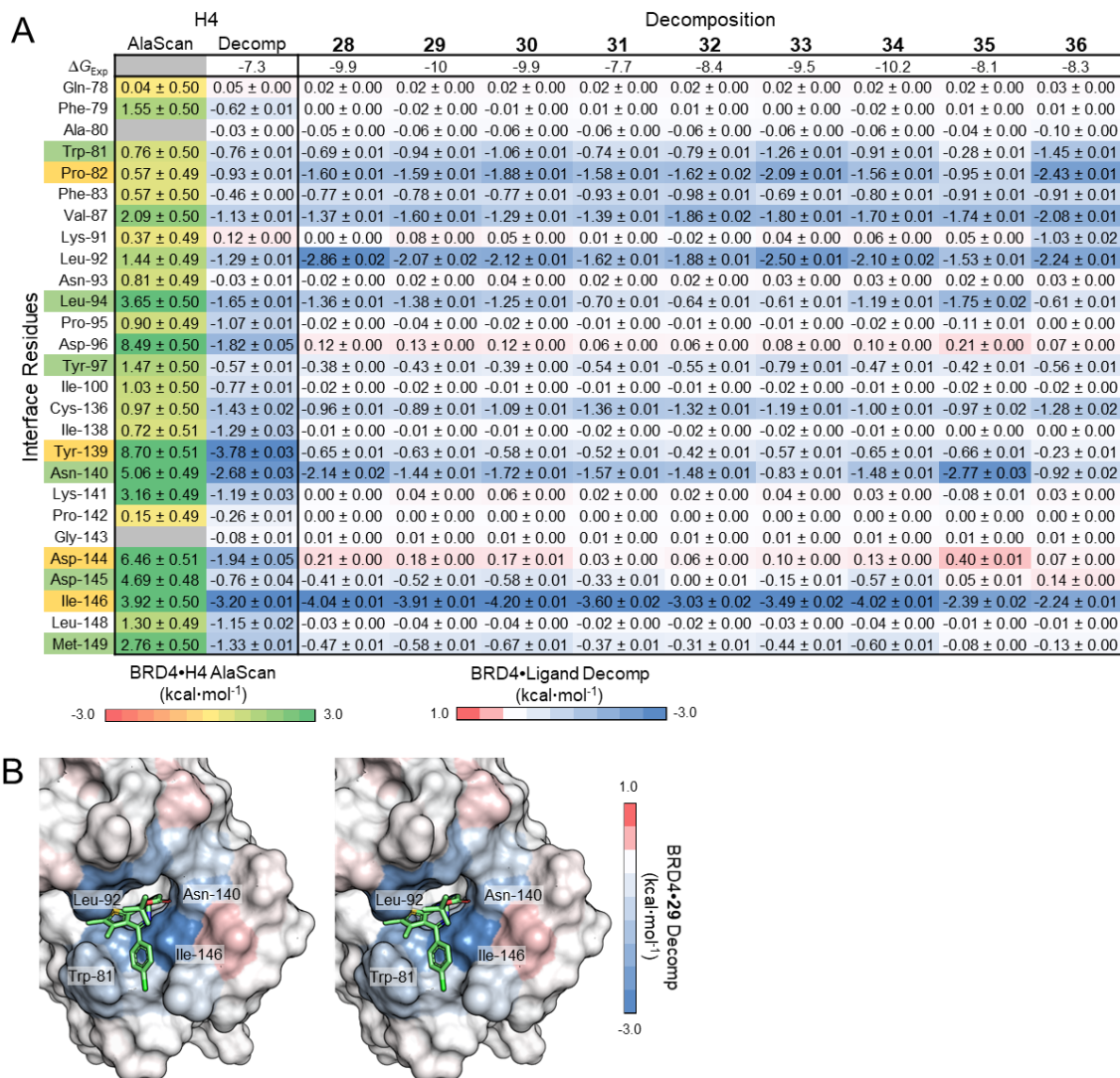


Figure 2.12. BRD4•H4 comparison with inhibitors. **(A)** Residues on BRD4 at the interface of the protein-protein interaction. On the left, residues are color-coded based on experimental mutagenesis studies. Known hot spots residues are highlighted in green. Mutations that do not greatly impact the interaction are highlighted in yellow. The first two columns show alanine scanning and per-residue decomposition at the specific residue in the protein-protein complex, respectively. The third to last column show the per-residue decomposition of at the specific residue for each co-crystallized inhibitor. Alanine scanning and per-residue decomposition energies are color-coded per the scales in **Fig. 2.11B** and **Fig. 2.11C**, respectively. Experimental ΔG are shown in the first row for each complex. Numbers are shown in kcal·mol⁻¹. **(B)** Surface representation of BRD4 colored by per-residue decomposition energy with compound **29**.

Compounds **28-36** do not appear to engage the hot spots of BRD4 as strongly as inhibitors of the other protein-protein interactions considered in this work (**Fig. 2.12A**). Compounds **28-32** and **34** share a common core structure and binding mode. One example is the interaction between **29** (JQ1) and BRD4, which consists primarily of hydrophobic contacts and van der Waals interactions (**Fig. 2.12B**). A triazole ring on **29** forms a hydrogen bond with the Asn-140 hot spot, mimicking Lys-5(ac). The chlorobenzene ring of the compound occupies the pocket formed by Lys-8(ac). The *t*-butyl acetate moiety in **29** is replaced by different moieties in the other analogs and extends out of the pocket into solution. In **28**, the phenyl substituent forms a π - π interaction with Leu-92, a non-hot-spot residue. Compound **34** differs by replacing the *t*-butyl group with a methyl, thereby exposing less of the compound into solution and reducing the compound's binding affinity by approximately 13 nM and IC₅₀ by more than 50 percent. Despite decreasing the size of this moiety, there is no observable effect on the energy decomposition of these two compounds. While the residue is more distant from Lys-8(ac) in the peptide, the residue serves to lodge the three fused rings of the compounds in the binding pocket.

2.2.9 Mimicking Hot Spots on the Protein Ligand. Following our extensive study of small-molecule binding to receptor hot spots, we wondered how effectively existing small-molecule protein-protein interaction inhibitors mimic the position of hot spots on the ligand protein of the complexes considered in this work. To explore compound overlap with protein ligand hot spots, we resorted to pharmacophore modeling. Hot-spot residues located on the ligand protein were identified from the literature for the Bcl-xL•Bak, MDM2•p53, XIAP•Smac, and BRD4•H4 complexes. For IL-2•IL-2R α , we could not identify a set of experimental hot-spot residues on the α -subunit; we selected all residues on IL-2R α at the interaction interface for the pharmacophore modeling. Pharmacophore hypotheses were generated to summarize the physiochemical properties of hot-spot residue on the protein ligand using Schrödinger Phase [115, 116].

The overlap between compounds and hot spots on the protein ligand is system specific. For example, moieties on the inhibitors of the MDM2•p53 and XIAP•Smac interactions showed the most significant overlap with protein ligand hot spots. Small-molecule inhibitors of Bcl-xL•Bak, IL-2•IL-2R α , and BRD4•H4 showed the lowest degree of overlap. For Bcl-xL•Bak, compounds overlap primarily with Leu-78 and Ile-85, while showing no detected overlap with the other four hot-spot residues, Val-74, Arg-76, Ile-81, and Asp-83 (**Fig. 2.13**). Compounds **2**, **3** and **6** have moieties that mimic the hydrophobic side chain of Leu-78 on the Bak peptide while compounds **1-4** mimic the hydrophobic moiety of Ile-85 on the Bak peptide. On the MDM2•p53 complex, small molecules showed excellent overlap with all three of the hot-spot residues on p53, namely Phe-19, Trp-23, and Leu-26 (**Fig. 2.14**). Compounds on MDM2•p53 antagonists mimic the indole of Trp-

23 with similar indole or benzenes rings. The aromatic ring of Phe-19 is generally occupied by hydrophobic aliphatic moieties on the compounds (**7**, **11**, **12**, **14**, **15**, and **16**), while aromatic groups on the compounds were introduced to mimic the hydrophobic side chain of Leu-26. Like MDM2•p53, small-molecule inhibitors of XIAP•Smac showed significant overlap with side chains of the Smac ligand. There are four hot spots residues at the N-terminal region of Smac (**Fig. 2.15**). Val-2, Pro-3, and Ile-4 all have hydrophobic pharmacophore features on their side chains. Most of the XIAP•Smac inhibitors we have considered in this study contain moieties that overlap with and mimic the side chains of these residues. The exceptions are compounds **18** and **23** overlapping with Val-2 and compound **22** overlapping with Ile-4. For IL-2•IL-2R α , there was remarkably little overlap between inhibitors of this interactions and hot spots located on IL-2R α . There are 10 residues on the α -subunit at the IL-2•IL-2R α interface (**Fig. 2.16**). All the compounds for this interaction have an amine group that mimics the positive charge on Arg-36. Finally, BRD4•H4 interaction antagonists mimicked the two acetylated lysine residues (**Fig. 2.17**). Compounds **32**, **33**, **35**, and **36** mimic the hydrophobic pharmacophore feature on Lys-5(ac), while compounds **28**, **29**, **33**, and **34** mimic the hydrophobic pharmacophore feature on Lys-8(ac).

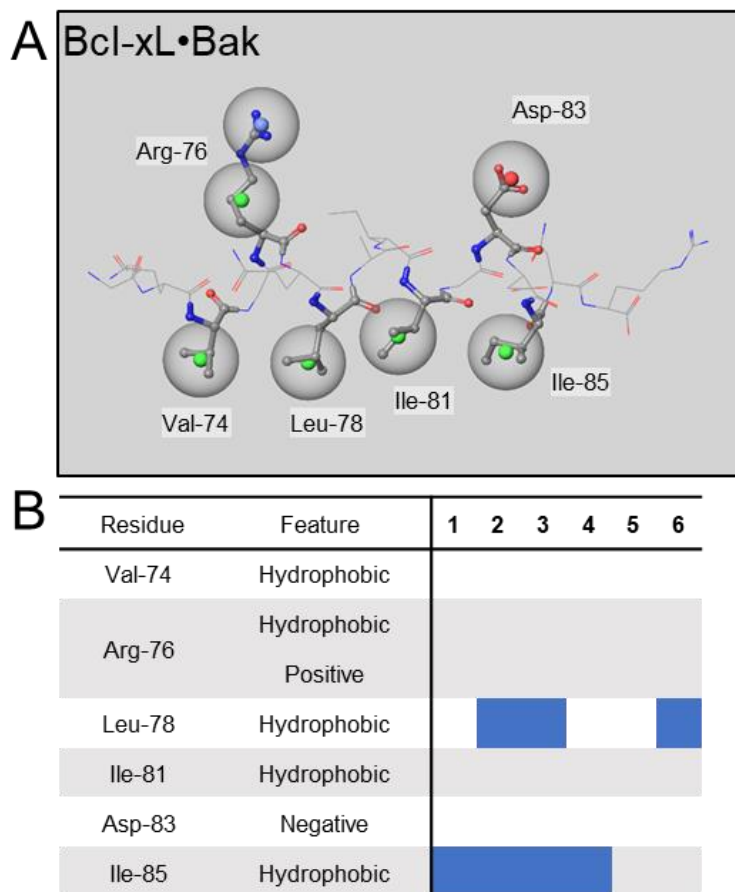


Figure 2.13. Hot-spot residues on the protein ligand Bak and overlap with inhibitors in Bcl-xL•Bak. **(A)** The pharmacophore model of the protein-protein interaction complex. The pharmacophore features for each protein ligand are shown as small colored spheres: Hydrophobic (H, green), positive charge (P, dark blue), and negative charge (N, red). Tolerances are shown in transparent gray spheres around the pharmacophore centers. **(B)** The hot-spot residues on the protein ligand are shown with associated pharmacophore features for that residue. For each compound (shown column-wise), if a chemical moiety matches the associated pharmacophore feature at that residue, the box is colored blue.

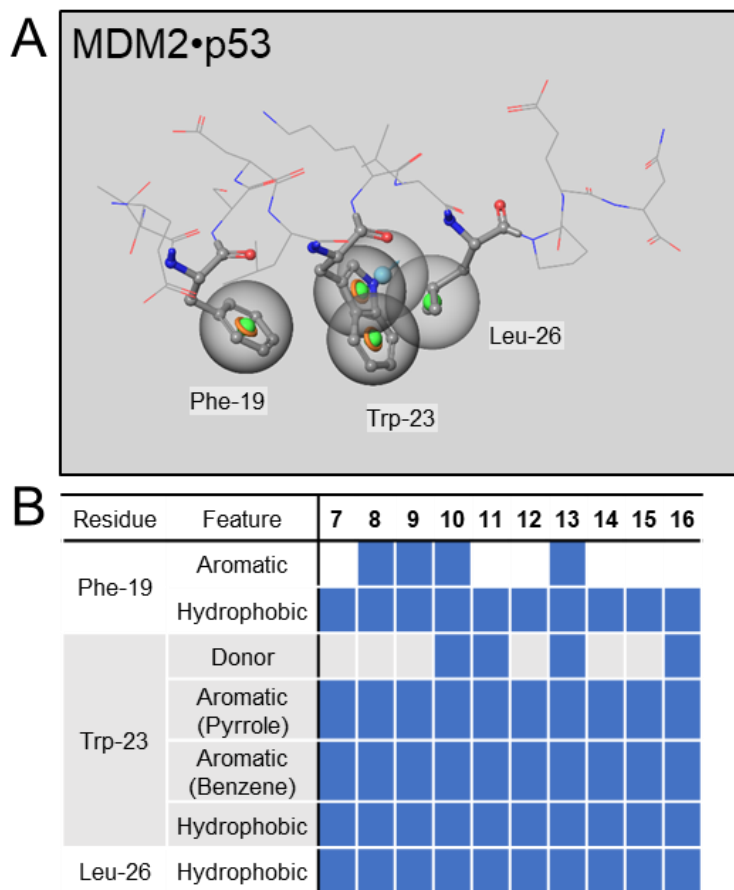


Figure 2.14. Hot-spot residues on the protein ligand p53 and overlap with inhibitors in MDM2•p53. (A) The pharmacophore model of the protein-protein interaction complex. The pharmacophore features for each protein ligand are shown as small colored spheres: Hydrogen bond donor (D, light blue), hydrophobic (H, green), and aromatic ring (R, tan). Tolerances are shown in transparent gray spheres around the pharmacophore centers. (B) The hot-spot residues on the protein ligand are shown with associated pharmacophore features for that residue. For each compound (shown column-wise), if a chemical moiety matches the associated pharmacophore feature at that residue, the box is colored blue.

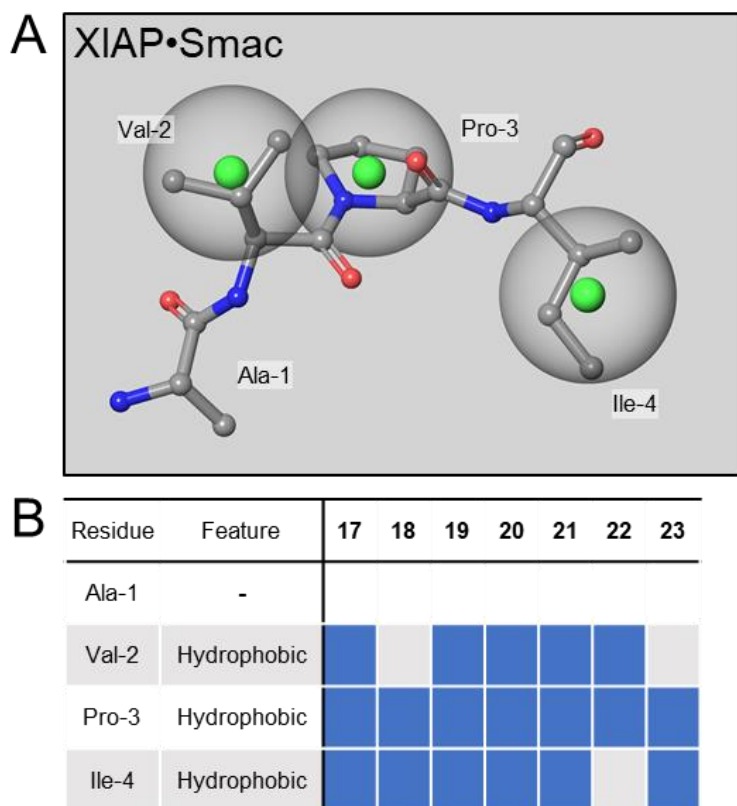


Figure 2.15. Hot-spot residues on the protein ligand Smac and overlap with inhibitors in XIAP•Smac. **(A)** The pharmacophore model of the protein-protein interaction complex. The pharmacophore features for each protein ligand are shown as small colored spheres: Hydrophobic (H, green). Tolerances are shown in transparent gray spheres around the pharmacophore centers. **(B)** The hot-spot residues on the protein ligand are shown with associated pharmacophore features for that residue. Alanine residues had no pharmacophore features that could be considered and were left blank. For each compound (shown column-wise), if a chemical moiety matches the associated pharmacophore feature at that residue, the box is colored blue.

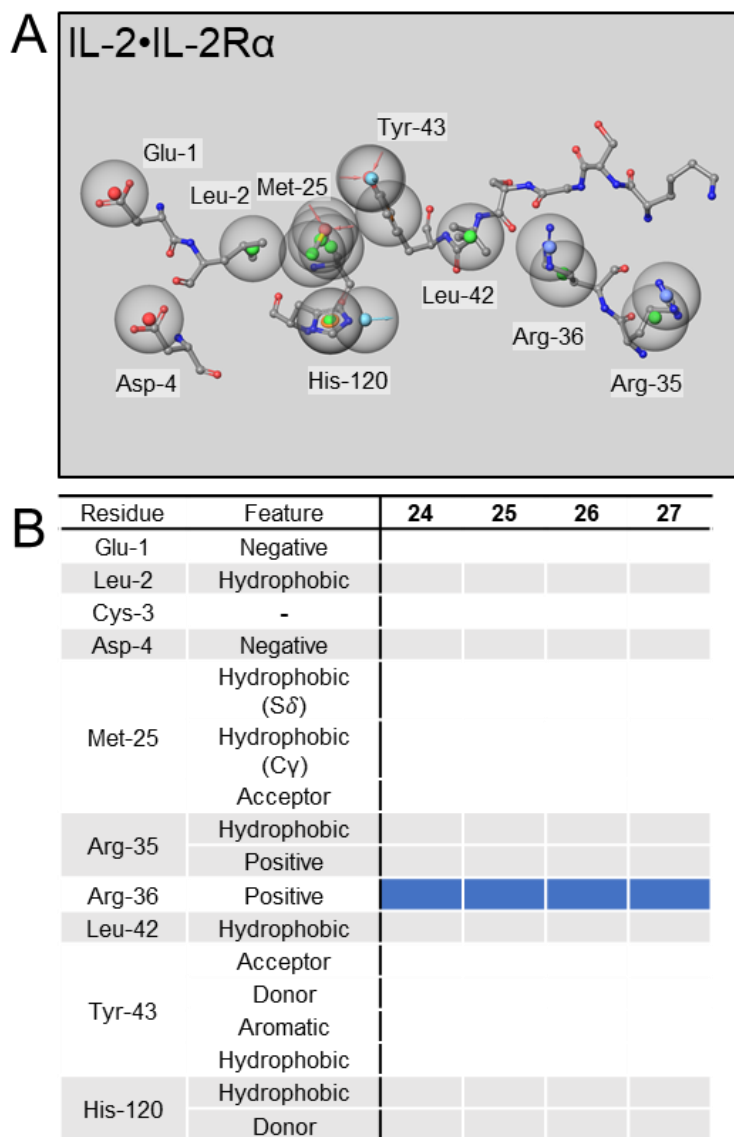


Figure 2.16. Hot-spot residues on the protein ligand IL-2R α and overlap with inhibitors in IL-2•IL-2R α . **(A)** The pharmacophore model of the protein-protein interaction complex. The pharmacophore features for each protein ligand are shown as small colored spheres: Hydrogen bond acceptor (A; red), hydrogen bond donor (D, light blue), hydrophobic (H, green), positive charge (P, dark blue), negative charge (N, red), and aromatic ring (R, tan). Tolerances are shown in transparent gray spheres around the pharmacophore centers. **(B)** The hot-spot residues on the protein ligand are shown with associated pharmacophore features for that residue. Cysteine residues had no pharmacophore features that could be considered and were left blank. For each compound (shown column-wise), if a chemical moiety matches the associated pharmacophore feature at that residue, the box is colored blue.

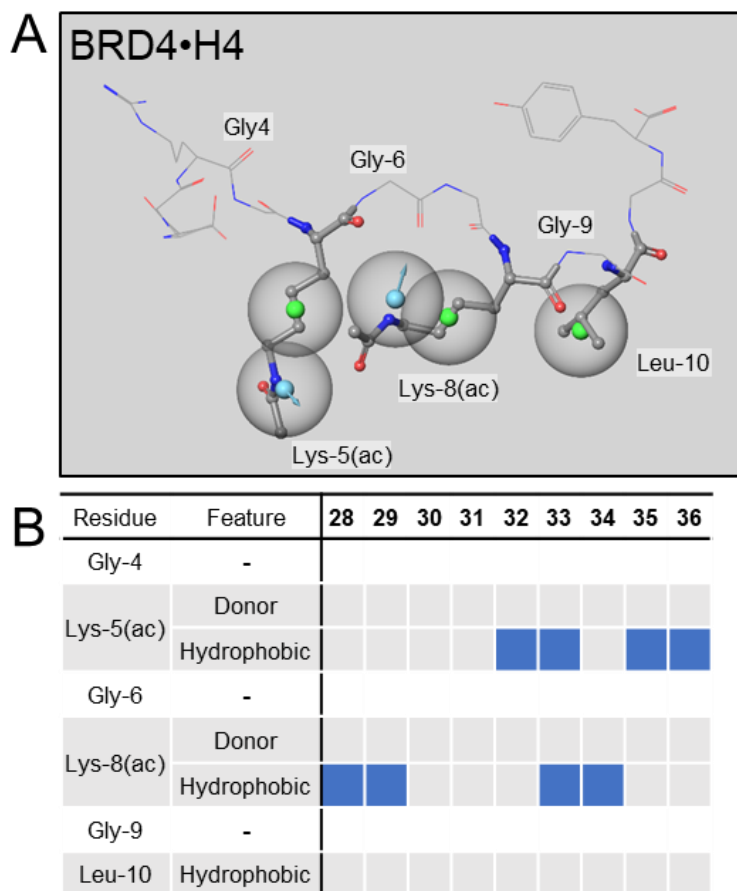


Figure 2.17. Hot-spot residues on the protein ligand H4 and overlap with inhibitors in BRD4•H4. **(A)** The pharmacophore model of the protein-protein interaction complex. The pharmacophore features for each protein ligand are shown as small colored spheres: Hydrogen bond donor (D, light blue) and hydrophobic (H, green). Tolerances are shown in transparent gray spheres around the pharmacophore centers. **(B)** The hot-spot residues on the protein ligand are shown with associated pharmacophore features for that residue. Glycine residues had no pharmacophore features that could be considered and were left blank. For each compound (shown column-wise), if a chemical moiety matches the associated pharmacophore feature at that residue, the box is colored blue.

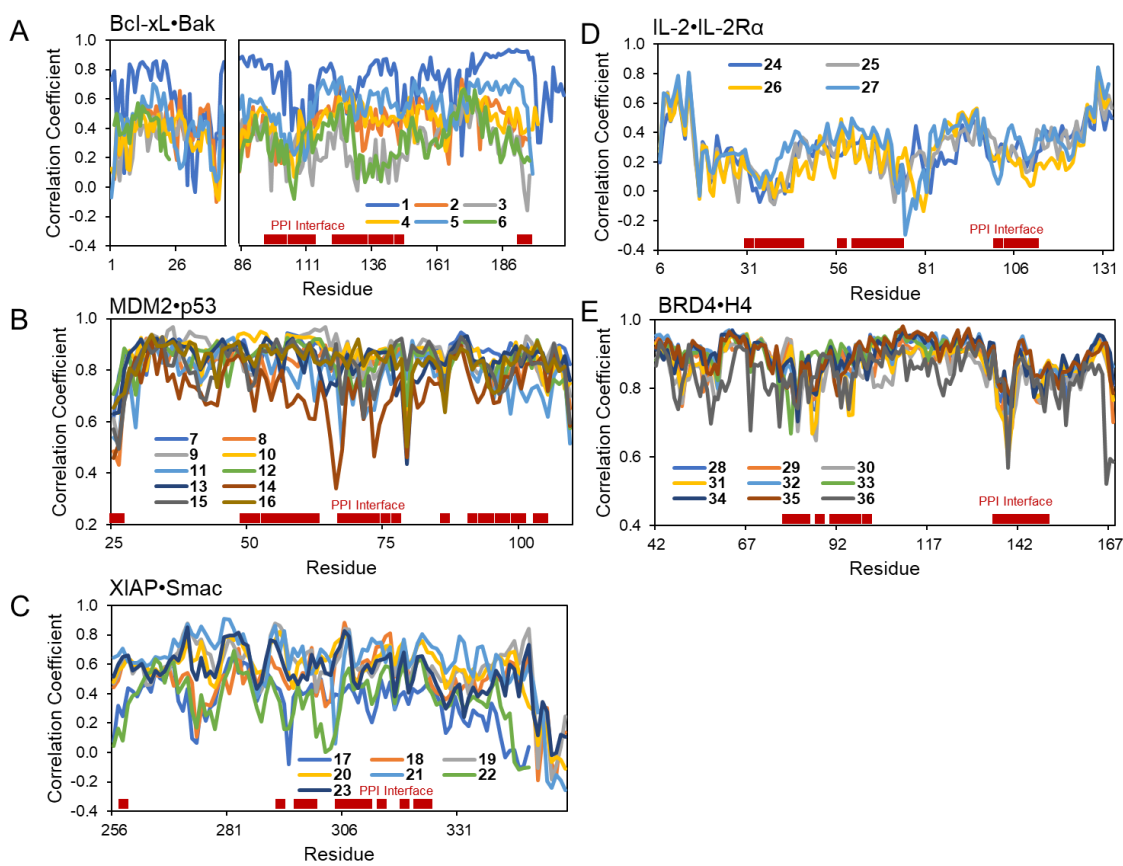


Figure 2.18. Similarity in the dynamics of inhibitors with the native ligand on the protein receptor. (A-E) Dynamic cross-correlation matrices (DCCM) were generated for the protein-protein and protein-inhibitor complexes. At each residue, the Pearson correlation coefficient between the cross-correlation of the protein-protein and protein-compound for (A) Bcl-xL•Bak, (B) MDM2•p53, (C) XIAP•Smac, (D) IL-2•IL-2R α , and (E) BRD4•H4. A positive correlation at a residue indicates that the residue on the protein receptor in the protein-protein and protein-compound complexes are moving in a similar manner, while a negative correlation indicates that the residue is moving in an opposite manner between the protein-protein and protein-compound complexes. Residues at the protein-protein interaction interface on the protein receptor are shown as red squares at the bottom of each panel.

2.2.10 Effect of Native Protein Ligand and Small-Molecule Inhibitors on Receptor

Dynamics. A question of interest is whether small-molecule inhibitors mimic the effect of the native ligand protein on the dynamics of the receptor. We compared the dynamics of the native ligand and each of the inhibitors using dynamic cross-correlation matrices (DCCM) [117]. A dynamic cross-correlated matrix measures the correlation of motion between each residue or ligand with every other residue or ligand in the complex. A correlation coefficient of 1 means that the residues are moving in the same direction, while -1 corresponds to two residues that are moving away from each other. We determine the similarity between the correlated motions between the native ligand and inhibitors for every residue on the protein receptor. Generally, the dynamical motion of residues on hot spots of the protein receptor is correlated between protein-protein and protein-compound complexes (**Fig. 2.18**).

The mean correlations are highest in MDM2•p53 and BRD4•H4, with mean correlations of 0.95 ± 0.04 and 0.87 ± 0.00 , respectively. These are the two systems that overlap with both the greatest number and percentage of hot spots on the ligand. These are followed by XIAP•Smac, Bcl-xL•Bak, IL-2•IL-2R α , with mean correlations of 0.50 ± 0.01 , 0.46 ± 0.01 , 0.31 ± 0.01 , respectively. Similarly, compounds of the XIAP•Smac and Bcl-xL•Bak interactions overlap with one or two hot spots in their respectively interactions, while compounds of IL-2•IL-2R α only overlap with a single hot spot. There is a significant difference in the correlations between residues at the interaction interfaces and residues outside of the interface (Bcl-xL•Bak, Mann-Whitney rank-sum test, $p = 1.55 \times 10^{-5}$; MDM2•p53, Mann-Whitney rank-sum test, $p = 4.81 \times 10^{-6}$; XIAP•Smac, Mann-Whitney rank-sum test, $p = 1.76 \times 10^{-6}$; IL-2•IL-2R α , Student's t -test, $p = 6.38 \times 10^{-15}$; BRD4•H4, Mann-Whitney rank-sum test, $p = 5.41 \times 10^{-29}$).

2.3 DISCUSSION

We carried out extensive molecular dynamics simulations followed by end-point free energy calculations of protein-protein and protein-compound complexes. The goal was to characterize how effectively existing small molecules that inhibit protein-protein interactions mimic the binding of the protein ligand to the receptor. We selected five protein-protein interactions for which small-molecule inhibitors have been developed and co-crystalized with their target. The five protein-protein interactions fall into three categories, namely primary (BRD4•H4 and XIAP•Smac), secondary (MDM2•p53 and Bcl-xL•Bak), and tertiary (IL-2•IL-2R α). A total of 36 compounds were considered. The compounds range in binding affinity and physicochemical properties. In most cases, the compound ligand efficiencies were below what is generally accepted as drug-like, namely 0.3. The only exceptions were the BRD4 antagonists. This is attributed to the

fact that large compounds had to be prepared to disrupt the protein-protein interactions, particularly for the tighter interactions.

In addition to computational alanine scanning, we carried out decomposition energy calculations. The interaction energy of compounds with individual amino acids is determined using a similar approach to MM-GBSA, except that only atoms on the ligand and a single residue are included. Unlike alanine scanning, these calculations do not require a mutation and therefore the effect on the interaction of the compound with the protein is not affected by the absence of the residue. One can imagine that mutation of an amino acid with a side chain that plays a critical role in the structural integrity of the protein could lead to changes to the stability of the protein that may introduce changes to the free energy of binding of a small molecule that are unrelated to the protein-compound interaction. While in general we see a good correlation between $\Delta\Delta G_{\text{MM-GBSA}}^{\text{AlaScan}}$ and $\Delta E_{\text{GBTOT}}^{\text{Decomp}}$, there were, for each case, several exceptions. Several residues were predicted to be hot spots, yet the intermolecular decomposition of the native peptide to the residue was not very strong. This includes Phe-146 for the Bcl-xL•Bak, Trp-310 for XIAP•Smac, and Phe-79, Val-87, Leu-94, Asp-96, Lys-141, and Asp-145 for BRD4•H4. The weaker decomposition energy interaction energy can be attributed to the fact that the contribution of these residues to binding involves entropy factors that are not considered in the decomposition energy calculations. Conversely, we found several examples of native protein ligands that strongly engaged residues that were not found to be hot spots in the alanine scan. Examples include Val-126 on Bcl-xL, Thr-26 and Lys-51 on MDM2, and Thr-308 on XIAP. Interaction of small molecules with residues that are not considered hot spots provides an opportunity to design compounds with greater specificity.

Computational alanine scanning and decomposition energy calculations were carried out for each protein-protein complex, and decomposition energies were calculated for each protein-protein and protein-compound complex. Overall inspection of these color-coded maps reveals that small molecules strongly engage a set of hot spots on the receptor. These include Phe-97, Leu-130, and Arg-139 in Bcl-xL•Bak, Leu-54 and Val-93 in MDM2•p53, Leu-307, Glu-314, and Trp-323 in XIAP•Smac, Phe-42, Lys-43, and Glu-62 in IL-2•IL-2R α , and Ile-146 in BRD4•H4. More notable, however, was the number of predicted hot spots that were not engaged by small-molecule inhibitors. These include Arg-100 and Glu-129 in Bcl-xL•Bak, Met-50 and Tyr-100 in MDM2•p53, Trp-310 and Tyr-324 in XIAP•Smac, Tyr-45 and Pro-65 in IL-2•IL-2R α , and Asp-98, Tyr-139, Lys-141, Asp-144, Asp-145, and Met-149 in BRD4•H4. The lack of engagement of these hot spots on Bcl-xL, MDM2, and BRD4 are examples of residues further outside the binding pocket, which may provide additional opportunities for the design of inhibitors.

In addition to engagement of receptor hot spots, we explore how effectively small-molecule protein-protein interaction inhibitors mimic hot spots on the protein ligand. We found that only the MDM2•p53 and XIAP•Smac compounds effectively mimicked the side chain of hot spots on the protein ligands of these interactions. Surprisingly, substantially less overlap between compounds and ligand hot spots was found for Bcl-xL•Bak, IL-2•IL-2R α , and BRD4•H4 interactions. The most surprising finding was that of IL-2•IL-2R α , which showed no overlap with any of the hydrophobic residues on IL-2R α . This is not unexpected considering that for Bcl-xL and IL-2 antagonists, a fragment-based approach was followed to develop small-molecule antagonists. The lead optimization efforts were mainly driven by an attempt to occupy the pockets at the protein interface and maximize interaction with receptor residues. This is evidenced by overall strong engagement of receptor hot spots by Bcl-xL and IL-2 inhibitors. Generally, it is not essential that small molecules engage all hot spots at the protein-protein interface or engage all hot spots to the same extent as the ligand protein to be effective inhibitors. However, small molecules that bind more tightly to hot spots than the native ligand may result in more potent inhibitors that will likely exhibit greater ligand efficiency.

Finally, few studies have explored how effectively small-molecule protein-protein interaction inhibitors mimic the dynamics of the native protein ligand. In a protein-protein complex, the binding of the ligand modulates the motion of a receptor. We quantify this effect using cross-correlated dynamical maps, which determine the correlation of motion between the ligand protein and every residue on the receptor. We compare the correlation coefficients between residues on the protein receptor with the native protein ligand and with compounds. Interestingly, we found that in some systems, such as MDM2•p53 and BRD4•H4, the protein ligand dynamics correlated remarkably well to small-molecule inhibitors of these interactions. For Bcl-xL•Bak and XIAP•Smac, correlation between native protein and compound was overall weak. Interestingly, very little correlation between protein ligand and compound dynamical changes on the receptor were found for IL-2•IL-2R α . Lead optimization efforts for the design of protein-protein interactions seldom consider the effect of compounds on the dynamics of the receptor. This approach may be used as a strategy to favor compounds that more closely mimic the dynamics of the native ligand.

2.4 MATERIALS AND METHODS

2.4.1 Structure Preparation. A set of protein structures corresponding to both protein-protein and protein-compound complexes were identified from 2P2I [118] and UniProt [119] (Table 2.1). In total, 36 protein-compound complexes for inhibitors of five protein-protein interactions that possess binding affinity data were collected. Existing experimental measure of

binding affinity (K_d), inhibition constants (K_i), and concentration at which 50% inhibition is observed (IC_{50}) for each complex were identified from the literature, whenever available. Protein-compound affinities were confirmed against data in the PDBbind [120], BindingMOAD [121], and BindingDB [122] databases, whenever possible.

The structure of each complex was retrieved and prepared using Protein Preparation Wizard in the Schrödinger software package (Schrödinger LLC, New York, NY, 2015). Bond orders were assigned, hydrogen atoms were added, and disulfide bonds were created. Water molecules were retained while additional ions and heteroatom groups aside from the inhibitor were discarded. Missing side chains and loops were introduced using the Prime module [123]. The resulting protein and compound structures were protonated at pH 7.0 using PROPKA [124] and Epik [125] in Schrödinger, respectively. Schrödinger concurrently samples sidechain orientations and protonation states to optimize hydrogen bonding, charge interactions, and orientations of hydroxyl, thiol, terminal amide groups of Asn, Gln, and His residues. Protein-protein complexes were separated into monomeric chains and protein-inhibitor complexes were separated into protein and compound structures for molecular dynamics simulations.

2.4.2 Molecular Dynamics Simulations. Prepared structures were used to run molecular dynamics simulations using the AMBER14 software package [126]. Each compound was assigned AM1-BCC [127] charges and gaff [128] atom types using *antechamber* [129]. Crystal water molecules were retained. Complexes were immersed in a box of TIP3P [130] water molecules. No atom on the complex was within 14 Å from any side of the box. The solvated box was further neutralized with Na^+ or Cl^- counterions using the *tleap* program. Simulations were carried out using the GPU accelerated version of the *pmemd* program with ff12SB [131] and gaff [128] force fields in periodic boundary conditions. All bonds involving hydrogen atoms were constrained by using the SHAKE algorithm [132], and a 2 fs time step was used in the simulation. The particle mesh Ewald [133] (PME) method was used to treat long-range electrostatics. Simulations were run at 298 K under 1 atm in NPT ensemble employing Langevin thermostat and Berendsen barostat. Water molecules were first energy-minimized and equilibrated by running a short simulation with the complex fixed using Cartesian restraints. This was followed by a series of energy minimizations in which the Cartesian restraints were gradually relaxed from 500 kcal·Å⁻² to 0 kcal·Å⁻², and the system was subsequently gradually heated to 298 K with a 48 ps molecular dynamics run. For each complex, we generated 10 independent simulations (replicates) that are each 10 ns in length. The initial velocities for each trajectory were randomly assigned using the built-in random number generator in Amber. This ensures that the trajectories follow different paths in phase space, resulting in better sampling of the molecular dynamics of the protein-protein or protein-ligand

complexes. In total, 100 ns of simulation were carried out for each protein-protein and protein-compound complex. We chose to run multiple short trajectories over one long trajectory as the former is widely accepted to result in more efficient sampling of the conformational space of the protein-protein or protein-ligand complex.

2.4.3 Free Energy Calculations. In each of the 10 trajectories (10 ns in length), the first 2 ns were discarded for equilibration. Snapshots were saved every 1 ps, yielding 8000 structures per trajectory. A total of 80000 snapshots were generated per 100 ns of simulation. 1000 snapshots were selected at regular intervals from the 80000 snapshots for free energy calculations using the *cpptraj* program [134]. The Molecular Mechanics-Generalized Born Surface Area (MM-GBSA) [135] method was used to calculate the free energy using the *MMPBSA.py* script [136]. The calculation using the GB method was performed with *sander* and Onufriev’s GB model [137, 138]. Solvent-accessible surface area (SASA) calculations were switched to the icosahedron (ICOSA) method, where surface areas are computed by recursively approximating a sphere around an atom, starting from an icosahedron. Salt concentration was set to 0.1 M. The entropy was determined by normal mode calculations [139] with the *mmpbsa_py_nabnmode* module by selecting 100 of the 1000 snapshots used in the free energy calculations at regular intervals. The maximum number of cycles of minimization was set to 10000. The convergence criterion for the energy gradient to stop minimization was 0.5. In total, 1000 frames were used for each MM-GBSA calculations while 100 frames were used for each normal mode analysis. All other parameters were left at default values.

The strength of protein ligand binding can be estimated by combining molecular-mechanics (MM) calculations with either Poisson-Boltzmann (PB) or generalized-Born (GB) surface area (SA) continuum solvation methods (MM-PBSA and MM-GBSA) [140]. One of the primary advantages of these methods is the modular nature of the individual contributions. For example, the MM-GBSA binding free energy is expressed as:

$$\Delta G_{\text{MM-GBSA}} = \Delta E_{\text{GBTOT}} - T\Delta S_{\text{NMODE}}$$

where ΔE_{GBTOT} is the combined internal and solvation energies, T is the temperature (298.15 K). ΔS_{NMODE} is the entropy determined by normal mode calculations. The total enthalpy from the generalized Born model, ΔE_{GBTOT} , is the sum of 4 components:

$$\Delta E_{\text{GBTOT}} = \Delta E_{\text{VDW}} + \Delta E_{\text{ELE}} + \Delta E_{\text{GB}} + \Delta E_{\text{SURF}}$$

where ΔE_{VDW} and ΔE_{ELE} are the van der Waals and electrostatic energies, respectively, and ΔE_{GB} and ΔE_{SURF} are the polar and non-polar desolvation energies, respectively. The total enthalpy solvation energy is determined using Generalized-Born (GB) solvation models (ΔE_{SOLV}):

$$\Delta E_{\text{GBTOT}} = \Delta E_{\text{GAS}} + \Delta E_{\text{SOLV}}$$

where ΔE_{SOLV} is the solvation free energy and ΔE_{GAS} is the molecular mechanical energies (gas-phase). The gas-phase energies are composed of two components:

$$\Delta E_{GAS} = \Delta E_{VDW} + \Delta E_{ELE}$$

The GB solvation free energy is expressed by the polar and non-polar contributions to the solvation free energy:

$$\Delta E_{SOLV} = \Delta E_{GB} + \Delta E_{SURF}$$

All binding energies are determined by:

$$\Delta E = E^{COM} - E^{REC} - E^{LIG}$$

where E^{COM} , E^{REC} and E^{LIG} are total energies corresponding to the complex, receptor, and ligand, respectively.

2.4.4 Alanine Scanning. Computational alanine scanning was calculated for each of the five protein-protein complexes using the *MMPBSA.py* script [40, 136]. The free energy change that results from mutation of a residue to alanine is averaged over all snapshots collected at equal intervals over the course of each trajectory. We carry out 10 runs \times 10 ns trajectories (100 ns total) for each complex, and we discard the first 2 ns of each run. We collect 100 snapshots at equal intervals (1000 snapshots in total). The 1000 frames collected for the MM-GBSA calculations were used for alanine scanning. Contact residues on the protein receptor within 5 Å of the protein ligand were identified in each protein-protein complex. Residues at the interface that are alanine and glycine cannot be mutated using the *MMPBSA.py* program and were not included. In addition, cysteine residues that form disulfide bonds were not included. Each remaining residue was mutated to alanine using Schrödinger Maestro and saved as a separate protein receptor and complex. Amber topologies were generated for the single mutant protein receptor and complex using *tLeap*. Alanine scanning was performed using the *MMPBSA.py* script as described above. The change in free energy between the original and mutated complexes ($\Delta\Delta G_{MM-GBSA}^{AlaScan}$) is calculated as:

$$\Delta\Delta G_{MM-GBSA}^{AlaScan} = \Delta G^{MUT} - \Delta G^{WT}$$

where ΔG^{MUT} is the complex with a single alanine mutation and ΔG^{WT} is the wild-type complex. A negative $\Delta\Delta G$ suggests that the mutation stabilizes the complex, while a positive value suggests that the mutation destabilizes it. Alanine scanning using *MMPBSA.py* assumes that a mutation will not change the overall dynamics of the complexed system, and merely changes the side chains of the lone mutated residue from the overall trajectory. We consider a residue to be a hot spot if $\Delta\Delta G_{MM-GBSA}$ is ≥ 1.5 kcal \cdot mol $^{-1}$. We chose this cutoff as a change of approximately 1.4 kcal \cdot mol $^{-1}$ in free energy (ΔG) results in an order of magnitude change in binding affinity (K_d).

2.4.5 Decomposition Energy. Per-residue MM-GBSA decomposition energies were calculated in addition to the total free energy through the *MMPBSA.py* script [40, 136] and the

above GB model. The decomposition scheme estimates the contributions to the total free energy on a per-residue basis ($idecomp = 2$). The per-residue decomposition energy includes only the contributions by the gas-phase and solvation energies ($\Delta E_{GBTOT}^{Decomp}$ and its components) and does not incorporate entropy (ΔS_{NMODE}).

Calculated free energies can be decomposed into specific residue contributions using either GB or PB implicit solvent models. These schemes were developed by Gohlke and co-workers [40]. The energy terms are decomposed using the following equation:

$$\Delta E^{Decomp} = \sum_{j \in COMPLEX \wedge j \in RECEPTOR} (\langle E^{COMPLEX}(i, j) \rangle - \langle E^{RECEPTOR}(i, j) \rangle) + \sum_{j \in COMPLEX \wedge j \in LIGAND} (\langle E^{COMPLEX}(i, j) \rangle - \langle E^{LIGAND}(i, j) \rangle)$$

where the first and second terms represent the average contribution over snapshots i from the MD simulation in residues j on the receptor and ligand, respectively. The term $E_{GBTOT}(i, j)$ corresponds to the contribution of the gas phase and solvation energies, that is:

$$E_{GBTOT}(i, j) = E_{GAS}(i, j) + E_{GBSOLV}(i, j) \\ = E_{VDW}(i, j) + E_{ELE}(i, j) + E_{GB}(i, j) + E_{SURF}(i, j)$$

where E_{VDW} and E_{ELE} are the van der Waals and electrostatic energies in the gas-phase (E_{GAS}), respectively. E_{GB} and E_{SURF} are the polar and non-polar contributions to the solvation free energy by the GB solvation model (E_{GBSOLV}), respectively. Entropy is not included in the decomposition method.

The popular GB^{OBC} model I [137] approximates the solvation electrostatic E_{GB} by an analytical formula:

$$E_{GB} = -\frac{1}{2} \sum_{ij} \frac{q_i q_j}{f^{GB}} \left(1 - \frac{\exp(-Kf^{GB})}{\epsilon} \right)$$

and

$$f^{GB} = \sqrt{r_{ij}^2 + R_i R_j \exp\left(-\frac{r_{ij}^2}{4R_i R_j}\right)}$$

where r_{ij} is the distance between atoms i and j , R_i and R_j are the effective Born radii of atoms i and j , K is the Debye-Hückel screening parameter, ϵ is the dielectric constant, and f^{GB} is a smooth function. Each atom in the GB model is represented as a sphere with radius ρ_i with charge q_i . The f^{GB} function is used to describe the distance between two atoms and their effective Born radii.

The non-polar contribution to the solvation free energy is calculated by approximating the total SASA of the molecule:

$$E_{SURF} = \gamma SASA + \beta$$

where γ and β are the surface tension and offset terms, respectively. By default, γ and β are 0.0072 and 0.0, respectively. The ICOSA method is used to determine SASA [40, 126]. In this method, surface areas are computed by recursively approximating a sphere around an atom. The first sphere is modeled as an icosahedron. In each subsequent step, the faces of the polyhedron are divided into four equal sized triangles to better approximate the sphere.

We consider a residue to be important to the interaction if it contributes ≤ -2.0 kcal \cdot mol $^{-1}$ in the decomposition energy. We found that this cutoff identified residues in the energy decomposition that were also identified from the computational alanine scan across each of the PPI systems.

2.4.6 Ligand Pharmacophore. A pharmacophore-based approach was used to identify how co-crystallized inhibitors overlapped with and mimicked known hot spots in each of the protein-protein complexes. First, the structures of each co-crystallized inhibitor were aligned to the reference protein-protein structure using the *align* function in PyMOL (version 1.8, Schrödinger, LLC). For each hot-spot residue, we defined a set of pharmacophore hypotheses corresponding to the physiochemical properties of the individual residue's side chain using the Phase package in Schrödinger [115, 116]. Phase has six built-in types of pharmacophore features: (i) hydrogen bond acceptor (A); (ii) hydrogen bond donor (D); (iii) hydrophobe (H); (iv) negative ionizable (N); (v) positive ionizable (P); and (vi) aromatic ring (R). By default, Phase does not place a hydrophobic feature for the aromatic ring feature; it will not identify an aromatic or aliphatic group in a lipophilic area of a binding pocket. Since hydrophobic moieties are commonly used to mimic aromatic rings, we generate a separate hydrophobic feature for aromatic ring pharmacophores. In the pharmacophore calculations, we use existing conformers from the aligned structures and set the intersite distance matching tolerance to 2.5 Å. All other parameters were set at default values.

2.4.7 Dynamic Cross-Correlation Matrix. Dynamic Cross-Correlation Matrices (DCCM) were calculated from the set of 1000 snapshots from the free energy calculations using the *matrix correl* function in the *cpptraj* program [134]. Correlation matrices were calculated by grouping together atoms of the same residues with no averaging by atom mass. Each DCCM is truncated to only include residues on the protein receptor. In this $n \times n$ matrix, each element is the correlation in dynamical motion between two residues in a protein-protein or protein-compound complex. Elements in a $1 \times n$ vector of this matrix thus corresponds to the correlation between a single residue with every other residue in the complex. A correlation coefficient of 1 corresponds to two residues moving in the same direction, while -1 corresponds to two residues that are moving

away from each other. To determine the similarity in motion between the native protein-protein and a protein-compound complex, at each residue, we calculate the Pearson's correlation coefficient between the two corresponding $1 \times n$ vectors for that residue.

2.4.8 Statistical Analysis. Values are expressed as mean \pm standard error, unless otherwise specified. Performance to rank-order complexes was evaluated using three correlation metrics, namely the Pearson's correlation coefficient (r), Spearman's rho (ρ), and Kendall's tau (τ). To test for significance between two groups, we first test for both normality within each group and equality of variances between groups using the Shapiro-Wilk test ($\alpha = 0.05$) and Levene's test ($\alpha = 0.05$), respectively. A choice of Student's t -test (normal distribution and equal variance), Welch's t -test (normal distribution and unequal variance), or Mann-Whitney rank-sum test (non-normal distribution) for significance is then selected as appropriate. Tests of statistical significance and correlation analysis were performed using the SciPy [141] package in Python.

Chapter 3

MIMICKING INTERMOLECULAR INTERACTIONS OF TIGHT PROTEIN-PROTEIN COMPLEXES FOR SMALL-MOLECULE ANTAGONISTS

3.1 INTRODUCTION

Protein-protein interactions range from weak ($K_d > 1000$ nM), moderate (100 nM $< K_d < 1000$ nM), to tight ($K_d < 100$ nM) [25-27]. Kastiris and co-workers found that 68% of a set of 144 curated co-crystallized protein-protein interactions were both tight and occurred over a large binding interface (> 1000 Å²) [142]. Yet, despite the gradual increase in the number of small-molecule protein-protein interaction inhibitors [49, 51, 58, 64, 143-146], only a handful among them are inhibitors of tight protein-protein interactions as shown in the previous chapter. Small-molecule inhibitors of tight interactions tend to be much larger than typical drugs, and generally have poor ligand binding efficiencies, which could explain the tendency for these compounds to fail in clinical trials. The development of small molecules that disrupt tight protein-protein interactions could expand the number of druggable proteins for the development of therapeutic agents.

Considering the ever-expanding size of commercial compound libraries, virtual screening could provide an avenue for developing chemical starting points that can be turned into potent inhibitors of tight protein-protein interactions. To the best of our knowledge, one study has used virtual screening to identify small-molecule inhibitors of a tight protein-protein interaction without extensive cycles of design and synthesis [147]. The most common approach for discovery of protein-protein inhibitors involves the identification of a weak-affinity fragment whose affinity is optimized by growing the fragment into neighboring pockets. For Bcl-xL•Bax [58] and IL-2•IL2-R α [48], fragment-based approaches and synthesis of derivatives to optimize binding to pockets at the protein-protein interfaces led to highly potent small-molecule inhibitors of the protein-protein interactions.

Historically, most rational approaches for the design of small-molecule inhibitors of protein-protein interactions have focused on mimicking the position of amino acids located on the protein ligand of a protein-protein interaction [42, 148, 149]. Several studies have used interface residues of the protein ligand of a protein-protein interaction to guide the design of small-molecule inhibitors in virtual screening and lead optimization [150-156]. The most common approach is based on pharmacophore modeling to enrich libraries for compounds that possessed substituents that not only adopted the same position as the amino acid side chain, but also possessed similar physicochemical properties to the side chain. This strategy has worked reasonably well, although

it is worth mentioning that there are no examples to date of small molecules that disrupt tight protein-protein interactions that emerged directly from virtual screening. Another strategy consists of finding molecules that bind directly to the receptor with the hope that these compounds will disrupt the protein-protein interaction. This strategy has never led to inhibitors of tight protein-protein interactions. This is attributed to the fact that mere binding to the receptor is not sufficient and critical residues, sometimes referred to as hot spots, must be engaged.

Here, we explore a new approach that is focused on mimicking the pairwise binding profile of the native protein ligand to the protein receptor of a tight protein-protein interaction. We use a fingerprint approach to represent the binding profile of the protein ligand. We use this fingerprint to screen commercial chemical libraries for small molecules that mimic the pairwise interaction profile of the native protein ligand. We also consider the strategy of combining the fingerprint approach to the standard pharmacophore method that identifies molecules that mimic the position of protein ligand amino acids. We use the tight uPAR•uPA protein-protein interaction as a platform to test these methods. We dock a library of commercially-available compounds to uPAR and rank compounds using the binding profile of the native ligand following three different methods. Compounds are tested for activity using fluorescence polarization and microtiter-based ELISA confirm disruption of the uPAR•uPA interaction. We also test for direct binding with microscale thermophoresis. All active hits are tested for thiol reactivity, redox activity, and stability. An analog-by-catalog procedure to explore structure-activity relationships led to the selection and testing of several derivatives for each hit compound. To the best of our knowledge, this is the first example of the use of structure-based virtual screening that leads to small-molecule inhibitors of a tight protein-protein interaction.

3.2 RESULTS

3.2.1 uPAR•uPA as a Platform to Test Rank-Ordering Methods. The urokinase receptor (uPAR, *PLAUR*) is a cell surface glycosylphosphatidylinositol (GPI)-anchored receptor that is part of an extensive network of protein-protein interactions. Its binding partners include the serine proteinase urokinase type plasminogen activator uPA (*PLAU*) [157] and the glycoprotein vitronectin (*VTN*) [158-160]. The uPAR•uPA interaction is characterized by a β -turn on the protein ligand uPA ensconced in a large interface on the protein receptor uPAR [161-164]. The interaction is mediated by a 25-residue growth factor-like domain (GFD), and residues from a kringle-like domain of uPA (**Fig. 3.1A**).

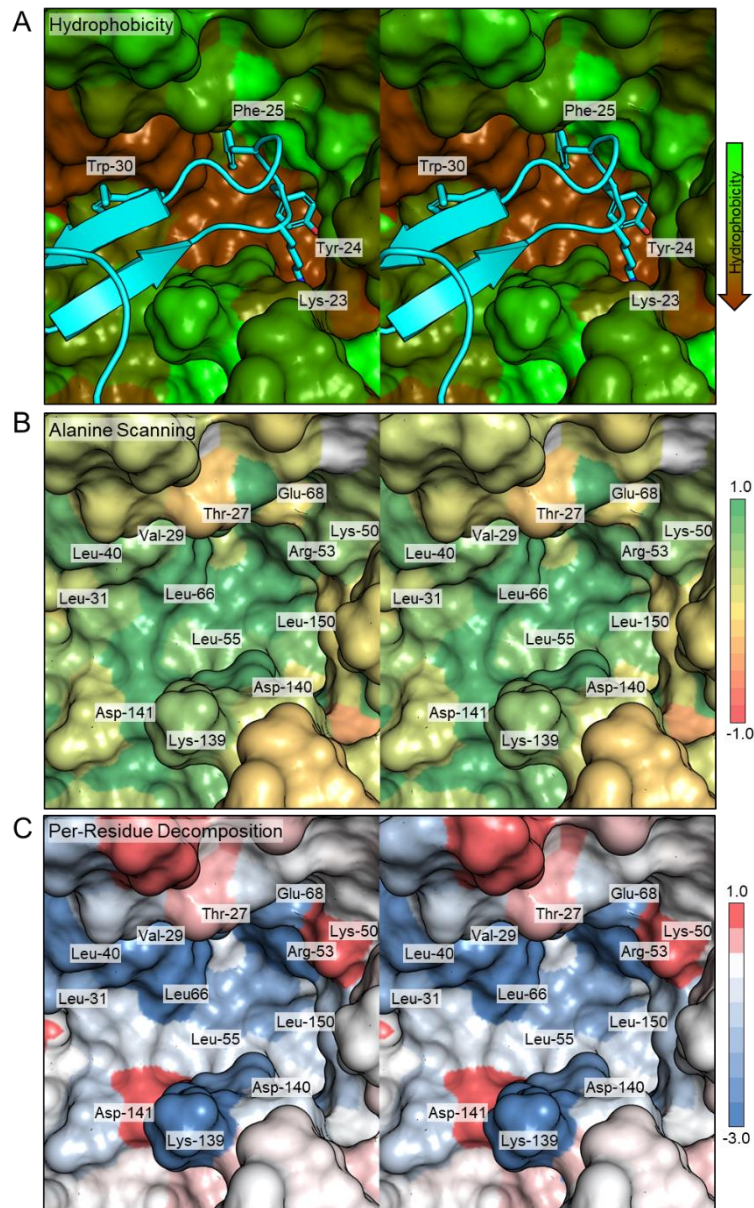


Figure 3.1. Structure of the uPAR•uPA binding pocket (PDB ID: 3BT1). (A) uPAR is shown in a surface representation with residues colored based on hydrophobicity. More hydrophobic residues are colored brown while more hydrophilic residues are colored green. uPA is colored cyan and shown in cartoon. The side chain of the four interface residues on uPA used in the pharmacophore analysis are shown in stick. (B) Experimental alanine scan of the uPAR•uPA binding pocket. The change in free energy between the mutated and wild-type complexes ($\Delta\Delta G$) after mutation of the residue to alanine is color-coded. (C) Per-residue decomposition energies of the uPAR•uPA binding pocket. The total enthalpic contribution of each residue is color-coded.

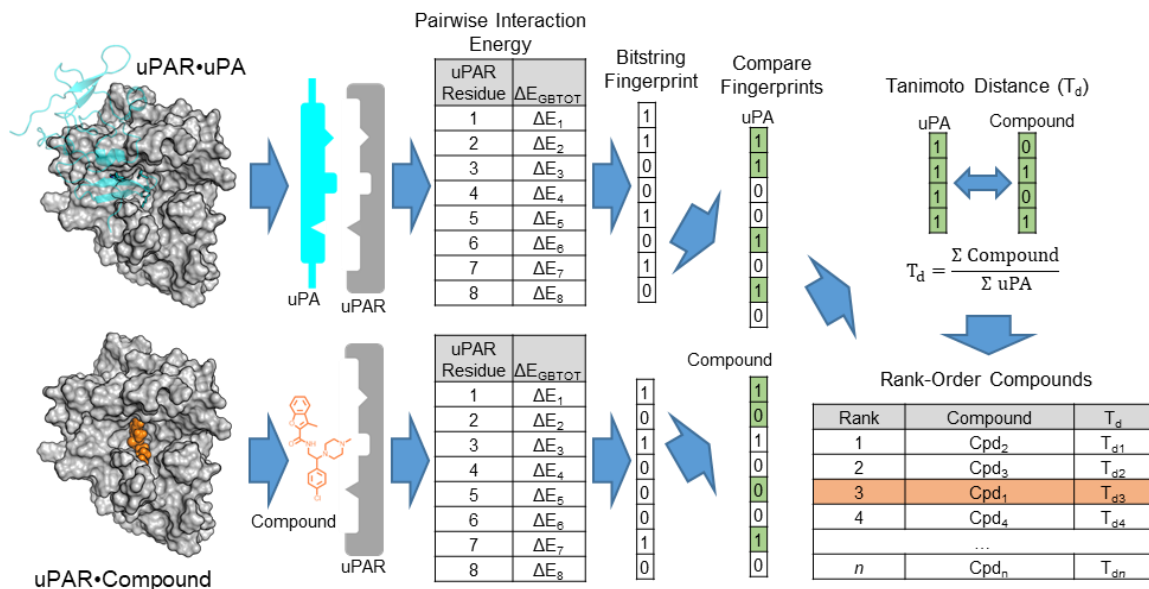


Figure 3.2. Workflow for the fingerprint method used to identify compounds that mimic the intermolecular binding interactions in the uPAR•uPA complex. The per-residue interaction energies of docked compounds are compared to those of the native protein ligand uPA. These interaction energies are used to generate a bitwise fingerprint, where each position on the fingerprint corresponds to the interaction energy between uPAR and the compound of interest. This fingerprint is compared to fingerprints of the native ligand uPA. Compounds are rank-ordered based on their Tanimoto distance with the fingerprints of uPA and total interaction energy ΔE_{GBTOT} .

Hot spot residues are featured on both uPAR and uPA, resulting in a tight ($K_d = 1$ nM) and stable ($k_{\text{off}} = 10^{-4} \cdot \text{s}^{-1}$) complex. In a comprehensive alanine scanning study, mutation at 15 residues on uPAR resulted in a significant decrease in binding affinity ($\Delta\Delta G \geq 1$ kcal $\cdot\text{mol}^{-1}$) [164]. Many of these residues are located in the binding pocket of uPA, including Leu-55, Tyr-57, Leu-66, and Leu-150. On uPA, the sidechain of five residues extend into the hydrophobic pocket of uPAR and are considered hot spots: Lys-23, Tyr-24, Phe-25, Ile-28, and Trp-30 [157].

3.2.2 A New Fingerprint Method to Rank-Order Compounds Based on their Ability to Mimic the Binding Profile of uPA to Residues on uPAR. Although previous studies have used key residues on the protein ligand to guide the design of small-molecule inhibitors, interactions with residues on the receptor have been generally ignored. Here, we utilize the interaction energies in the native protein-protein interaction to select top candidates that emerge from virtual screening of chemical libraries. We introduce a new approach that uses the native ligand to identify compounds that mimic the protein ligand's interaction with key receptor interface residues. To accomplish this, we use a fingerprint method summarized in **Fig. 3.2**. These fingerprints consist of strings of bits with length equal to the number of residues on the protein target, in our case uPAR. Each bit in the fingerprint corresponds to the interaction energy between the compound and a residue on uPAR. If the interaction energy between the ligand and the residue is greater than a threshold, the value of the bit is assigned to '1'. For compounds, the interaction energy consists of the computational decomposition energy. A value of '1' is assigned to a bit if the total decomposition energy ($\Delta E_{\text{Residue}}$) is less than -1.0 kcal $\cdot\text{mol}^{-1}$. For the native protein ligand, uPA, we generate two types of fingerprints based on either experimental data or computational decomposition energy. The first type of fingerprint is constructed using the experimentally-determined alanine scanning data of the uPAR•uPA complex (**Fig. 3.1B**). In this fingerprint, a value of '1' is assigned to a bit if the change in free energy following mutation of the residue to alanine ($\Delta\Delta G_{\text{AlaScan}}$) is greater than 1.0 kcal $\cdot\text{mol}^{-1}$. The second fingerprint is constructed using the decomposition energies from the molecular dynamics simulation of the uPAR•uPA complex (**Fig. 3.1C**). A value of '1' is assigned to a bit if the total decomposition energy ($\Delta E_{\text{Residue}}$) is less than -1.0 kcal $\cdot\text{mol}^{-1}$. If the threshold is not met, the value of the bit is '0'.

Following the docking of small molecules from commercial libraries to uPAR, a fingerprint is generated for each protein-compound structure. Each of these compound fingerprints is compared to the native protein ligand uPA fingerprint. Compounds with the most similar fingerprints to the protein ligand uPA are given higher priority. We use the Tanimoto distance (T_d) to compare the similarity between compounds and protein ligand uPA fingerprint. T_d is defined as the ratio between the number of bits in the fingerprint where both uPA and the compound have a

value of '1' over the number of '1' bits in the uPA fingerprint. The fingerprint generated from either alanine scanning or energy decomposition only includes positions where the corresponding uPA fingerprint has a value of '1'. However, the Tanimoto distance does not consider the positions of the specific bits when used to rank-order compounds. Similarly, the limited length of each fingerprint results in compounds sharing similar Tanimoto distances. If the Tanimoto distance of two compounds is equal, the total enthalpy from the MM-GBSA calculation of the compound (ΔE_{GBTOT}) is used to give higher priority to the compound with higher predicted binding affinity.

3.2.3 Application of the Fingerprint Method to Rank-Order Compounds using uPAR Interface Residues. We use the uPAR•uPA interaction as a platform to test our fingerprint method to rank-order compounds based on their interaction with receptor interface residues. The positions of the residues at the uPAR•uPA interface that were used to generate fingerprints are shown in **Fig. 3.3A**. Four residues on uPAR are present in both the uPA fingerprints based on the experimental alanine scanning and the fingerprints from energy decomposition: Leu-55, Leu-66, Leu-150, and His-166.

We separately rank-order the 5.1 million docked compounds based on their T_d value using (i) the uPA alanine scanning fingerprint and (ii) the uPA decomposition energy fingerprint. We select the top 500 candidates from each type of fingerprint. We examined how these compounds bind to each of these residues on uPAR (**Fig. 3.3B**). For compounds identified using the uPA fingerprint derived from energy decomposition, over 90% of the selected compounds interact favorably with Arg-53, Leu-55, Leu-66, Leu-150, and Ala-255. His-166 and Asp-25 interact with 41% and 86% of the compounds, respectively. For compounds identified using the uPA fingerprint derived from the experimental alanine scanning experimental data, 95% of the selected compounds interact with Leu-55, Leu-66, Leu-150, and His-166. Only 36% interacted with Arg-53, a residue that was included in the decomposition but not the alanine scanning fingerprint.

The top 500 compounds from both the energy decomposition and alanine scanning search strategies were independently clustered to 50 compounds using hierarchical clustering. Among the 50 compounds from each strategy, 29 from the uPA alanine scanning fingerprint and 24 from the uPA decomposition energy fingerprint were purchased for experimental validation. These 53 compounds were initially tested for binding to uPAR using a fluorescence polarization (FP) assay that we have previously developed (**Fig. 3.3C**) [165]. The assay consists of a fluorescently labeled α -helical peptide (AE147-FAM) that binds to uPAR at the uPAR•uPA interface.

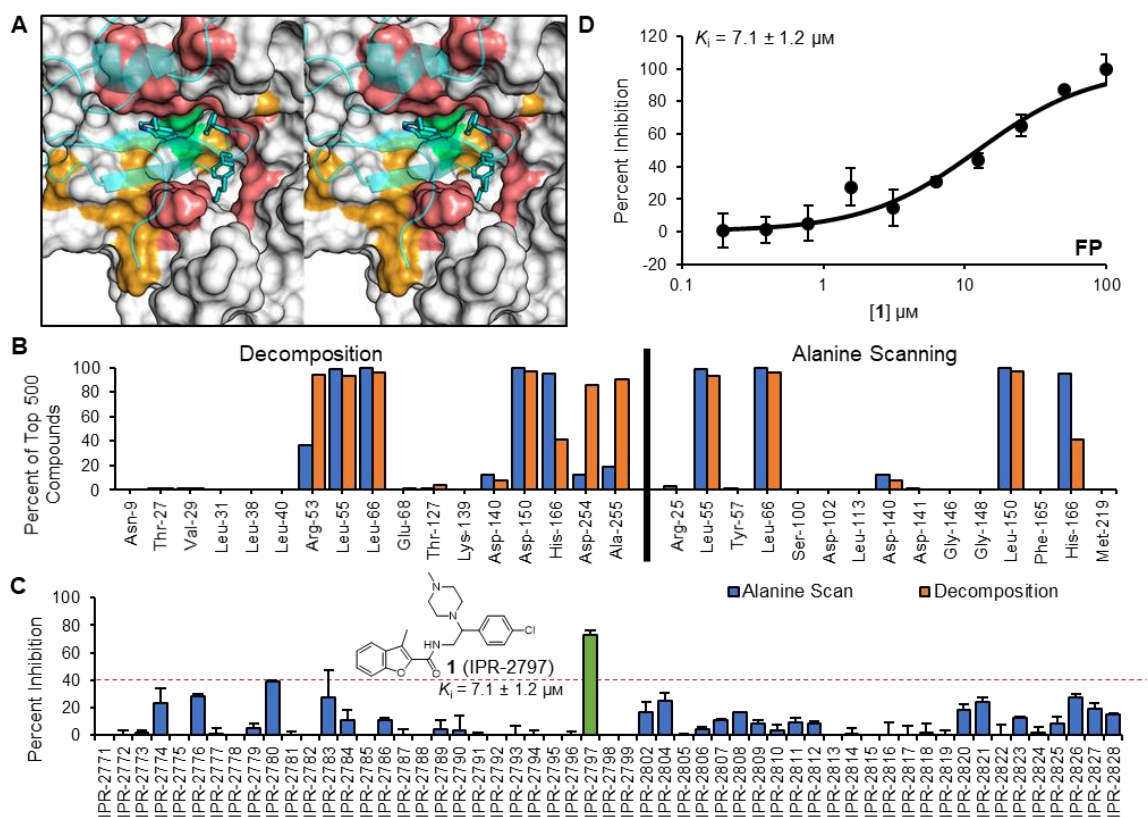


Figure 3.3. A virtual screen utilizing the interface residues of uPAR and validation of hits. **(A)** Residues used in the uPAR fingerprints are colored on the surface of uPAR as follows: (i) Experimental alanine scan (orange), (ii) decomposition (pink), (iii) both (green). uPA is transparently overlaid in cartoon, with the side chain of interface residues in stick. **(B)** Among the top-ranking 500 compounds from each of the fingerprints generated from decomposition energies or experimental alanine scanning, the proportion of compounds that overlap with each fingerprint residue. **(C)** Single-concentration FP screen of compounds resulting from the virtual screen based on uPAR residues. Each compound was screened in duplicate at 50 μM concentration (mean \pm SD). Hit compound **1** (IPR-2797) is highlighted in green. **(D)** Concentration-dependent FP assay measuring the inhibition of uPAR•AE147-FAM peptide interaction by **1** (IPR-2797). Representative of at least two independent experiments, where each concentration point is measured in duplicates (mean \pm SD).

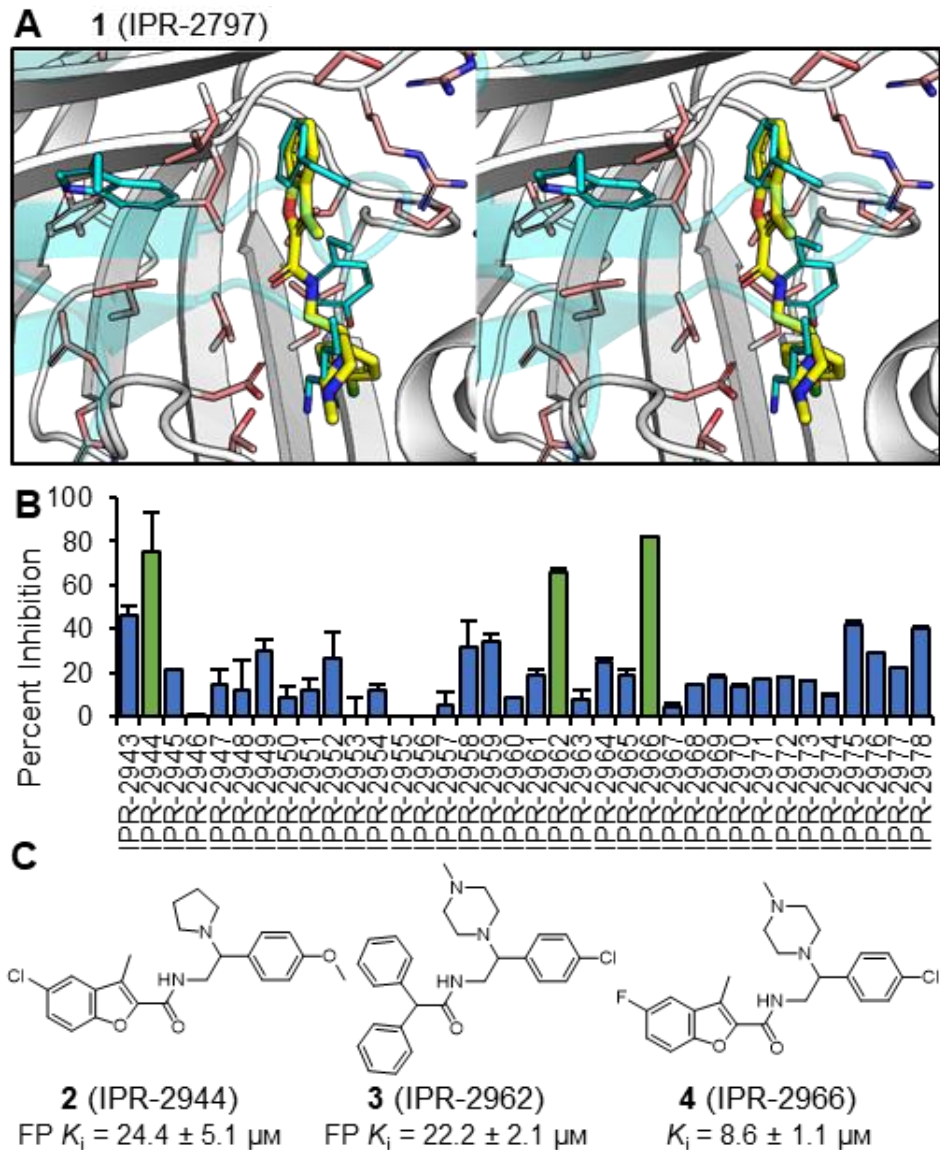


Figure 3.4. Screening the derivatives of compound **1** (IPR-2797). (A) The binding mode of **1** in the uPAR•uPA binding pocket. The compound is shown in yellow. uPAR is shown in white cartoon, with the side chain of interface residues shown in pink stick. uPA is shown in partial transparent cyan cartoon. The side chain of four interface residues on uPA are shown in stick and colored cyan. (B) Derivatives of **1** were screened at a single 50 μM concentration via the uPAR•AE147-FAM peptide FP assay in duplicates (mean \pm SD). Further pursued hits are highlighted in green. (C) Chemical structures of the pursued derivative hits.

One compound, **1** (IPR-2797), inhibited by more than 40%. A concentration-dependent study led to a K_i of $7.1 \pm 1.2 \mu\text{M}$ (**Fig. 3.3D**). A follow-up study using a microtiter ELISA method to analyze the compound inhibition of the uPAR•uPA_{ATF} interaction was performed. The compound did not show activity in the ELISA even at 100 μM .

We next assessed **1** for both reactivity and stability. The potential for **1** to covalently react with cysteine residues of a protein was evaluated using a (*E*)-2-(4-mercaptostyryl)-1,3,3-trimethyl-3H-indol-1-ium (MSTI)-based assay [166]. The compound did not react with MSTI suggesting that it is not thiol reactive. The compound was tested for redox activity by a Horseradish Peroxidase-Phenol Red (HRP-PR) assay and was found to be redox inactive at 100 μM concentration. Compound stability was tested in methanol, phosphate-buffered saline (PBS), and in the presence of uPAR by high-performance liquid chromatography-mass spectrometry (HPLC-MS). The compound showed the same retention time in HPLC and the mass remained the same, indicating that the compound is stable.

The predicted binding mode of **1** shows that the benzene of the benzofuran moiety overlaps with Phe-25 on uPA (**Fig. 3.4A**). In addition, a nitrogen in the piperazine ring of the compound is located near the positively charged amine on the side chain of Lys-23. Starting with the structure of **1**, we searched commercially-available libraries for analogs to conduct a preliminary structure-activity relationship (SAR) study. A set of 36 derivatives of **1** were identified, purchased, and screened at 50 μM using our FP assay (**Fig. 3.4B**). Three compounds, **2** (IPR-2944), **3** (IPR-2962), and **4** (IPR-2966), showed 75%, 66%, and 82% inhibition, respectively (**Fig. 3.4C**). The compounds inhibited the uPAR•AE147-FAM interaction in a concentration-dependent manner, although all were weaker than the parent compound. The most potent derivative, **4** ($K_i = 8.6 \pm 1.1 \mu\text{M}$), contains a fluorine atom on the aromatic ring of its benzofuran moiety.

3.2.4 Selecting Rank-Ordered Compounds using uPA Interface Residues. We explore another ranking method that strictly uses the interface residues located on the ligand protein to guide the selection of compounds. Our hypothesis is that small molecules docked to uPAR that mimic the side chain position of interface residues on the protein ligand will disrupt the uPAR•uPA interaction. At the uPAR•uPA interface, the side chain of five interface residues on uPA extend into the hydrophobic pocket of uPAR: Lys-23, Tyr-24, Phe-25, Ile-28, and Trp-30 (**Fig. 3.5A**).

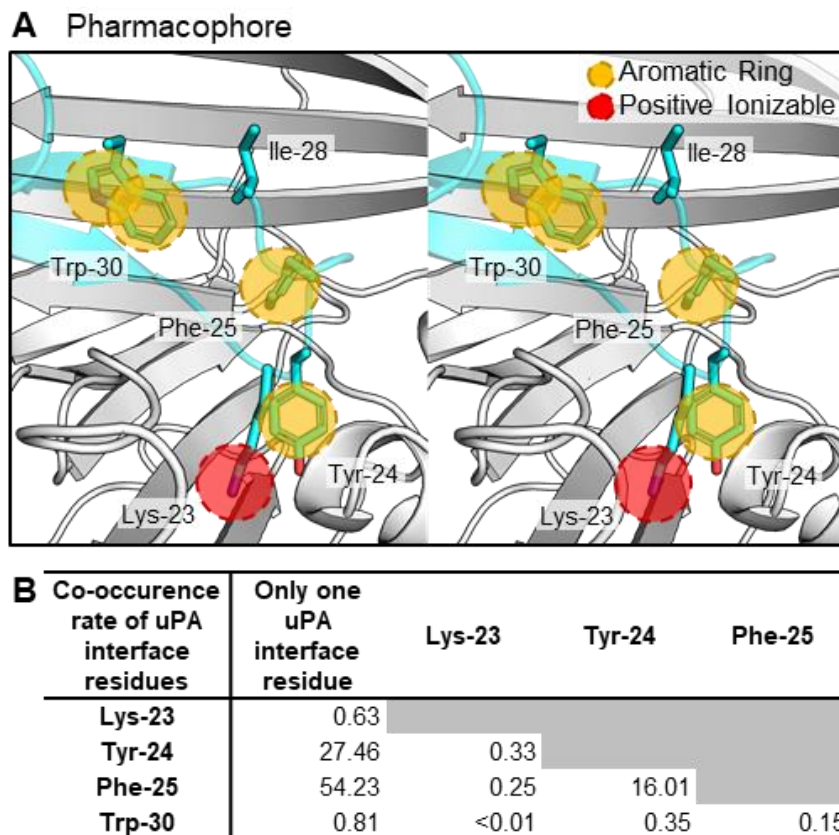


Figure 3.5. A virtual screen utilizing four interface residues of uPA. (A) Features of the pharmacophore model used to identify compounds that overlap with and mimic the interface residues of uPA. uPAR is shown in the background colored in white and shown in cartoon. uPA is shown in transparent cyan cartoon, with the five interface residues shown in stick. A pharmacophore model was used to assign features to four of the five interface residues (Ile-28 was excluded). (B) Co-occurrence of interface residues among all compounds that overlapped with at least one residue on uPA.

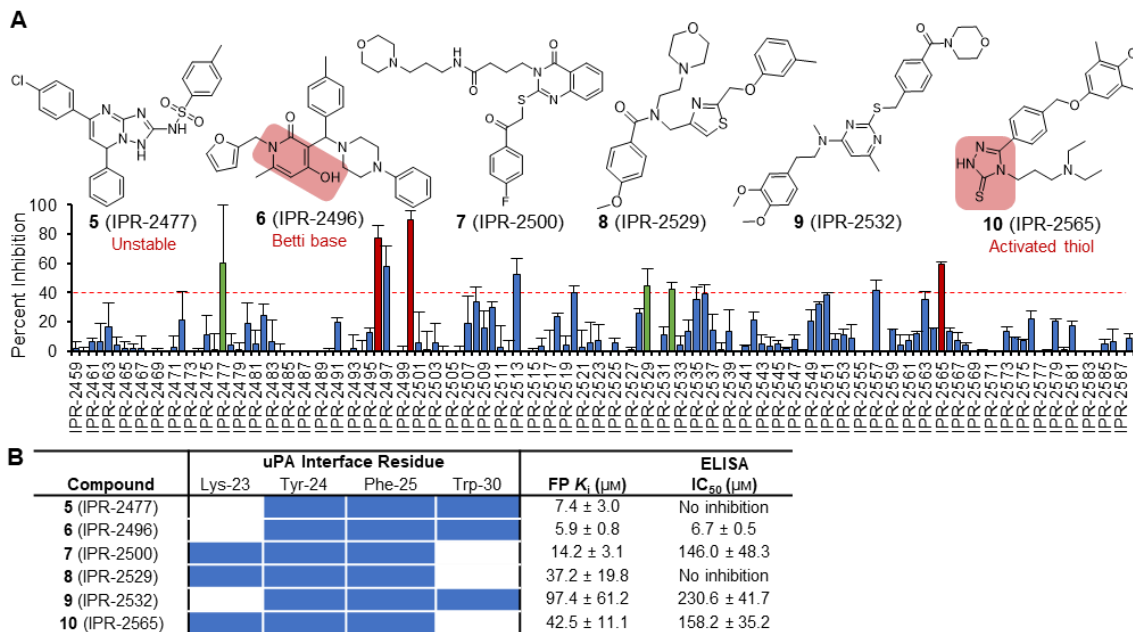


Figure 3.6. Validation of hits of virtual screen utilizing four interface residues of uPA. **(A)** Single-concentration FP screen of compounds resulting from the virtual screen based on uPA interface residues. Each compound was screened in duplicate at 50 μM concentration (mean \pm SD). Hits that are followed up are highlighted in green while those with problematic moieties are highlighted in red. Chemical structures of the highlighted molecules are shown above. **(B)** Overlap between the predicted binding mode of the hit molecules and the uPA residues are highlighted. FP and microtiter ELISA assays were used to measure the K_i and IC₅₀ of the compounds in inhibiting uPAR•AE147-FAM peptide and uPAR•uPA_{ATF} interactions, respectively. Results are based on at least two independent concentration-dependent experiments where each concentration point was measured in duplicates (mean \pm SD).

We use a pharmacophore approach [115, 116] to identify compounds with substituents that occupy the same position as the side chains of these residues. This approach consists of searching for small molecules that possess substituents that overlap with similar moieties (pharmacophores) on the side chain of amino acids. For example, a compound that possesses a benzene group that occupies the same position as the aromatic ring (pharmacophore) of a tyrosine residue is expected to disrupt binding of the residue to uPAR. In the pharmacophore model, the ϵ -amine on Lys-23 was modeled using a positive charge, while the benzene rings of Tyr-24 and Phe-25 were modeled using aromatic rings. We assigned separate aromatic ring features to the benzene and pyrrole rings on the indole of Trp-30.

For each of the 5.1 million docked compounds to uPAR, we determined whether there was an overlap with the defined pharmacophores on uPA. This resulted in 21312, 809846, 1297014, and 23047 matches for Lys-23, Tyr-24, Phe-25, and Trp-30, respectively. In total, approximately 1.8 million of the 5.1 million docked compounds overlapped with at least one of the pharmacophores corresponding to an interface residue on uPA (**Fig. 3.5B**). Among compounds that overlapped with a single residue, 54% and 27% matched the pharmacophores of Phe-25 and Tyr-24, respectively. Less than 2% of compounds overlapped with either the Lys-23 or Trp-30 pharmacophore. In contrast, 16% of compounds overlapped with both the Tyr-24 and Phe-25 pharmacophores.

We identified 1899 compounds that overlapped with 3 of these 4 residues, and no compounds that overlapped with all four residues. These compounds were hierarchically clustered to 200 using atom triplet Daylight fingerprints. We identified 130 commercially-available compounds that were purchased. These compounds were tested for binding to uPAR using our FP assay (**Fig. 3.6A**).

We initially tested the 130 compounds for activity using our FP assay at 50 μ M. We selected six compounds (**5-10**) that inhibited more than 40% (**Fig. 3.6B**). Compounds **5** (IPR-2477), **6** (IPR-2496), and **9** (IPR-2532) overlapped with Tyr-24, Phe-25, and Trp-30, while **7** (IPR-2500), **8** (IPR-2529), and **10** (IPR-2565) overlapped with Lys-23, Tyr-24, and Phe-25 (**Fig. 3.6B**). A concentration-dependent study for these compounds led to K_i values that ranged from 6 to 97 μ M. A follow-up study using a microtiter ELISA to analyze the compound inhibition of the uPAR•uPA_{ATF} interaction was performed. Although the ELISA cannot be used to obtain inhibition constants, it is a useful assay to determine whether compounds bind and disrupt the protein-protein interaction between uPA and uPAR. Four compounds, namely **6**, **7**, **9**, and **10** showed activity in the ELISA assay with IC_{50} s ranging from 7 to 230 μ M.

We assessed the reactivity and stability of all hit compounds. Compound **6** contains a Betti base that may cause the compound to be unstable and reactive, while **10** was thought to have a potential activated thiol group. The MSTI thiol reactivity assay was performed for each of the hits. Compound **10** readily reacted with MSTI as evidenced by a decrease in the fluorescence of MSTI. Compound **6** displayed no detectable MSTI reactivity, but we suspected that this was due to the unstable nature of the compound. The HRP-PR redox activity assay showed no significant redox capacity for compounds **5-10**. At this point, we decided to pursue compounds **5, 6, 8, and 9**. HPLC-MS stability assay for **8** and **9** showed the same single peak for both methanol and PBS buffers, indicating that the compounds were stable. However, **5** was found to be a mixture in HPLC-MS, suggesting that the compound inhibits in a non-specific manner. Compound **6** was pursued with reservation considering the reactive Betti base and found to bind non-specifically. Although compounds **8** and **9** inhibited uPAR, follow-up studies with derivatives showed that the effect was weak.

3.2.5 Selecting Rank-Ordered Compounds using both uPA and uPAR Interface Residues. We wondered whether combining our fingerprint method with the pharmacophore approach could yield small-molecule uPAR•uPA inhibitors. We combined the two search methods to identify a set of 69 compounds that overlapped with three of the four interface residues on uPA as well as engage residues in the uPA binding pocket on uPAR. A set of 39 compounds selected from among the 69 compounds were purchased for binding studies. The 39 compounds were tested in our FP assay (**Fig. 3.7A**). Seven compounds (**23-29**) were assessed using a concentration-dependent manner to determine the K_i values (**Fig. 3.7B**). Compounds **23** (IPR-2986), **26** (IPR-2992), **27** (IPR-2993), **28** (IPR-3089), and **29** (IPR-3193) overlapped with Lys-23, Tyr-24, and Phe-25 on uPA, while **24** (IPR-2987) and **25** (IPR-2989) overlapped with Tyr-24, Phe-25, and Trp-30. In comparison to hits that emerged from fingerprint or pharmacophore methods, the compounds had better K_i values that ranged from 6 to 52 μM . Only **25, 26, 28, and 29** inhibited uPAR•uPA_{ATF} based on a concentration-dependent study using our ELISA. The compounds had IC_{50} values of 68.3 ± 11.0 , 140.6 ± 19.0 , 172.8 ± 42.5 , and 24.8 ± 2.2 μM in the ELISA, respectively. The presence of α,β -unsaturated carbonyls on **23, 24** and **27** suggested potential reactivity with residue on uPAR and uPA. However, none of the hits from this screen showed reactivity with the activated thiol of the MSTI compound suggesting that the activity of the compounds is unlikely due to covalent bond formation. Compounds **24** and **27** displayed slight redox capacity in the HRP-PR assay, while compounds **23, 25, 26, 28, and 29** showed no redox activity. We focused our attention on **26, 28, and 29** as these compounds do not contain any problematic moiety and showed no covalent reaction

or redox activity. HPLC-MS analysis of these three compounds showed the compounds to be stable in both methanol and PBS buffers, with similar retention times.

Compound **26** binds into the uPAR•uPA pocket, mimicking the side chains of uPA and engaging interface residues on uPAR (**Fig. 3.8A**). A benzene moiety overlaps with Phe-25 on uPA. In the binding mode, a morpholino group is located between Lys-23 and Tyr-24 of uPA. In addition, a methyl substituent on the core quinoline ring points towards Trp-30. The core structure of **26** was used to identify derivatives. The derivatives we identified showed modifications at five substituents on **26** (**Fig. 3.8B**). A set of 136 derivatives of **26** were purchased and screened at 50 μM (**Fig. 3.8C**). The best hits were tested in concentration-dependent manner and their K_i s ranged from 2 to 37 μM (**Table 3.1**). The compounds were further tested in the uPAR•uPA_{ATF} ELISA assay to determine whether they can inhibit the protein-protein interaction. Only **32** (IPR-3026) and **39** (IPR-3116) failed to inhibit in the ELISA.

The best derivative among the **26** derivatives was **30** (IPR-3011). The binding mode of **30** shows that the additional moiety fits into a pocket lined by Asn-157, His-166, Leu-168, and Ala-255 on uPAR (**Fig. 3.8A**). The FP and ELISA inhibition curves of **26** and **30** are shown in **Fig. 3.9A** and **Fig. 3.9B**, respectively. The K_i and IC_{50} in the FP and ELISA assays for **30** are 2.5 ± 0.3 and 15.5 ± 1.4 μM , respectively. Compounds **26** (**Fig. 3.9C**) and **30** (**Fig. 3.9D**) were tested using microscale thermophoresis to assess direct binding to uPAR. The resulting K_d of **26** and **30** towards uPAR were 5.8 ± 1.3 and 2.0 ± 0.4 μM , respectively, consistent with the FP data for these compounds. Compound **30** and several of the other **26** derivatives have limited solubility. Like its parent **26**, **30** displayed no significant redox activity at 100 μM .

The binding mode of **28** shows overlap with Lys-23, Tyr-24, and Phe-25 on uPA (**Fig. 3.10A**). A set of 59 derivatives of **28** were purchased and screened by single concentration FP at 50 μM (**Fig. 3.10B**). The six best compounds, **44-49** (**Fig. 3.10C**), were tested in a concentration-dependent manner using the FP assay. While the analogs had K_i values ranging from 5 to 160 μM , the compounds did not inhibit in the ELISA. The first set of analogs, **44-46**, modify both the biphenyl and methyl group on the benzimidazole of the parent compound, yielding compounds that were less potent than the parent compound. The second set of analogs, **47-49**, modifies only the methyl group on the benzimidazole. Compound **49** (IPR-3485) lacks the methyl group entirely, resulting in threefold weaker inhibition constant than the parent. Compounds **47** and **48** possess aromatic substituents instead of the methyl group, resulting in K_i s of 5 μM in the FP assay. The lack of inhibition in the ELISA suggests that the compounds, while robust, may not be engaging the right set of residues to disrupt the full protein-protein interaction.

The binding mode of **29** reveals overlap with Lys-23, Tyr-24, and Phe-25 on uPA (**Fig. 3.11A**). A set of 20 derivatives of **29** were purchased and screened at 50 μM (**Fig. 3.11B**). The five best compounds, **50-54** (**Fig. 3.11C**), were tested at multiple concentrations using FP and uPAR•uPA_{ATF} ELISA. All but **53** (IPR-3235) (FP $K_i = 5.6 \pm 0.6 \mu\text{M}$, ELISA $\text{IC}_{50} = 52.0 \pm 4.1 \mu\text{M}$) and **54** (IPR-3236) (FP showed no inhibition; ELISA $\text{IC}_{50} = 65.3 \pm 12.3 \mu\text{M}$) had limited solubility in the two assays. Compound **29** was tested for direct binding in the MST assay and binds to uPAR with a K_d of $22.7 \pm 11.5 \mu\text{M}$.

3.3 DISCUSSION

The design of small-molecule inhibitors of protein-protein interactions has primarily focused on developing small molecules that bear substituents that mimic the position of amino acid side chains of the protein ligand in a protein-protein interaction. Here, we complement this approach by exploring a strategy that searches for small molecules that mimic the binding profile of the native protein ligand to the receptor of a protein-protein interaction. We hypothesize that small molecules that mimic the interaction of the native protein ligand to receptor are more likely to disrupt tight protein-protein interactions. To test this hypothesis, we introduced a quantitative approach to enable the comparison of the binding profiles of compounds to the binding profile of the protein ligand. We use a bitwise fingerprint to represent the pairwise interactions with amino acids on the receptor. When the ligand (protein or compound) engages a residue above a threshold, we assign the bit as '1'. The pairwise binding is based on the decomposition energy method that was introduced by Gohlke and co-workers to study the effect of individual amino acids on a protein-protein interaction [40]. The decomposition energy consists of the intermolecular energy between the ligand and each amino acid. This energy includes van der Waals, electrostatic, and polar and non-polar solvation energies. Following the creation of a fingerprint for the native protein ligand, in our case uPA, we used structure-based virtual screening to identify small molecules that shared a similar fingerprint. To do this, we docked a large number of compounds to uPAR, and generated fingerprints for all these compounds using the predicted binding pose. Compounds were ranked based on how closely their fingerprint matched the native ligand's (e.g. uPA). To accomplish this, we borrow from the cheminformatics field and use the Tanimoto distance to quantify the similarity between fingerprints.

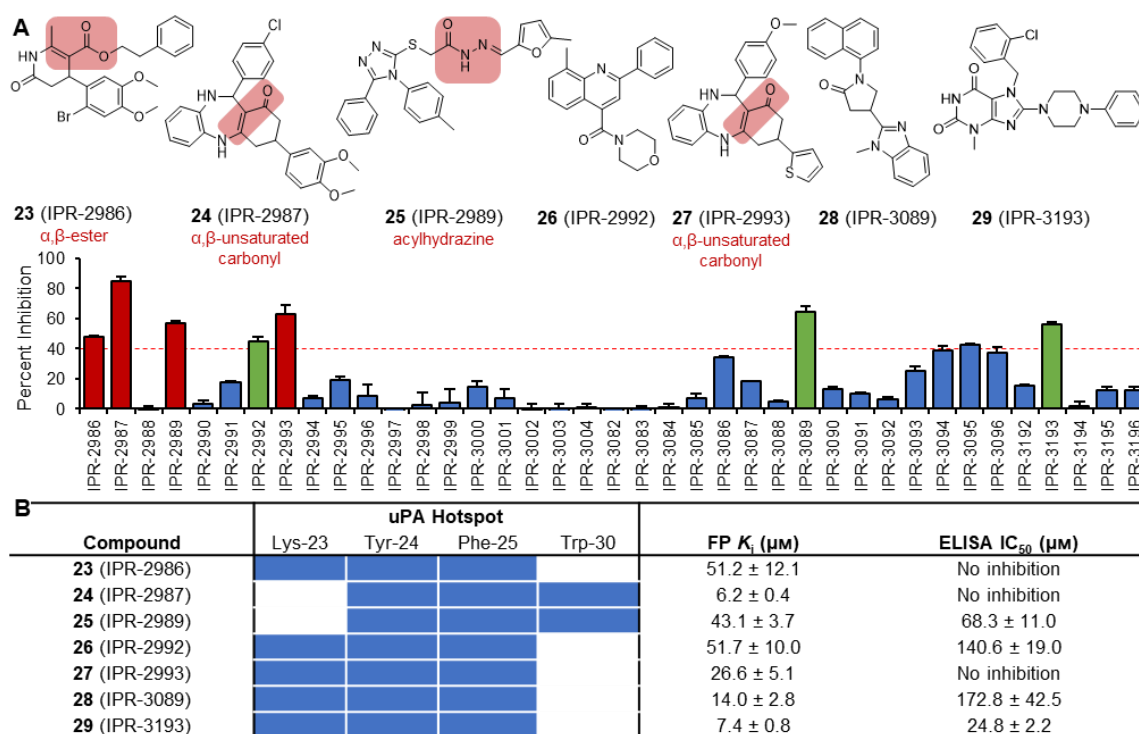


Figure 3.7. A virtual screen utilizing interface residues on both uPAR and uPA. **(A)** Single-concentration FP screen of compounds resulting from the virtual screen based on uPA interface residues. Each compound was screened in duplicate at 50 μM concentration (mean \pm SD). Hits that are pursued are highlighted in green while those with problematic moieties are highlighted in red. Chemical structures of the highlighted molecules are shown above. **(B)** Overlap between the predicted binding mode of the hit molecules and the uPA interface residues (IPR) are highlighted. FP and microtiter ELISA assays were used to measure the K_i and IC_{50} of the compounds in inhibiting uPAR•AE147-FAM peptide and uPAR•uPA_{ATF} interactions, respectively. Results are based on at least two independent concentration-dependent experiments where each concentration point was measured in duplicates (mean \pm SD).

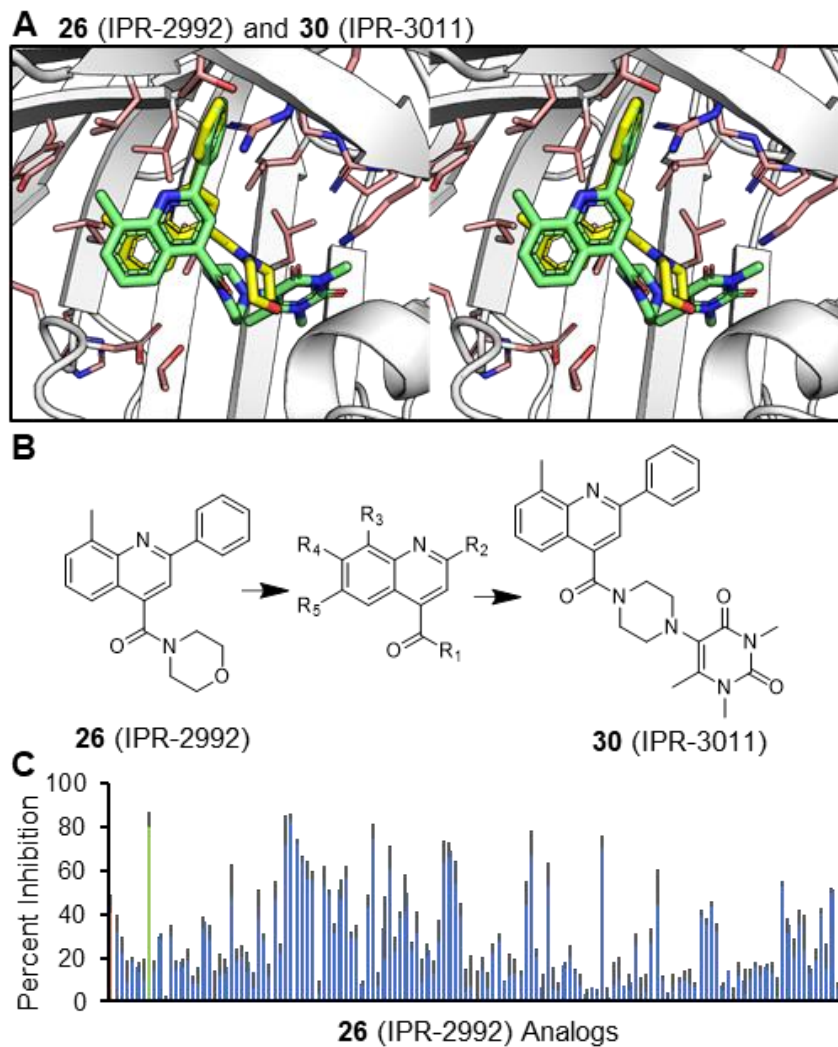
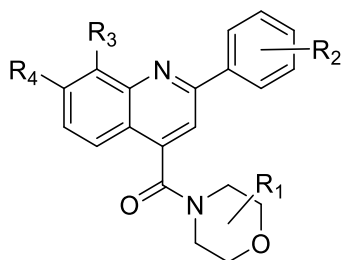
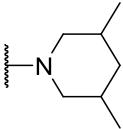
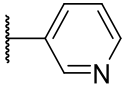
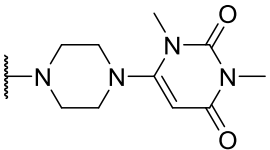
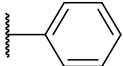
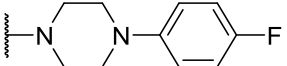
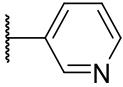
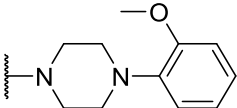
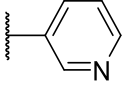
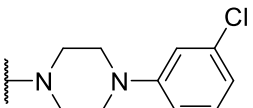
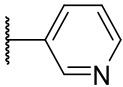
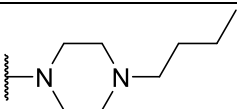
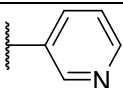
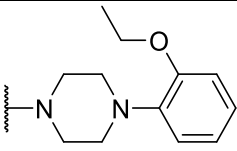
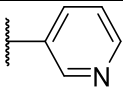


Figure 3.8. Testing the derivatives of **26** (IPR-2992) leads to **30** (IPR-3011). (A) The binding mode of **26** (IPR-2992) and **30** (IPR-3011) in the uPAR•uPA binding pocket. The binding mode of **30** (green) is overlaid on the binding mode of **26** (yellow). The additional ring at R₁ allows **30** to bind deeper in the uPAR•uPA pocket. uPAR is shown in white cartoon, with the side chain of interface residues shown in pink stick. (B) The core of **26** was used to identify analogs at 5 positions. Among the analogs discovered was **30** (IPR-3011). (C) Derivatives of **26** were screened at a single 50 μM concentration via the uPAR•AE147-FAM peptide FP assay in duplicates (mean ± SD). The parent compound **26** is highlighted in orange, while compound **30** is highlighted in green.

Table 3.1. Profiles of analogs of compound **26** (IPR-2992).



Compound	R ₁	R ₂	R ₃	R ₄	FP K _i (μ M) ^a	ELISA IC ₅₀ (μ M) ^a
30 (IPR-3011)			Me	H	2.5 \pm 0.3	15.5 \pm 1.4
31 (IPR-3015)			Me	H	37.1 \pm 1.9	171.5 \pm 35.7
32 (IPR-3026)			Me	Cl	15.4 \pm 3.4	No inhibition
33 (IPR-3036)			Me	H	5.2 \pm 0.5	62.7 \pm 10.7
34 (IPR-3037)			Me	H	8.0 \pm 1.1	99.5 \pm 14.2
35 (IPR-3038)			Me	H	11.9 \pm 2.0	67.9 \pm 11.2
36 (IPR-3039)			Me	H	8.4 \pm 0.6	37.9 \pm 2.4

37 (IPR-3040)			Me	H	26.6 ± 7.8	112.2 ± 8.6
38 (IPR-3103)			Me	H	5.2 ± 1.1	31.0 ± 2.1
39 (IPR-3116)			H	H	10.4 ± 3.1	No inhibition
40 (IPR-3117)			H	H	9.2 ± 1.9	331.6 ± 197.0
41 (IPR-3121)			H	H	3.4 ± 0.5	34.4 ± 7.4
42 (IPR-3134)			Me	H	9.9 ± 1.4	122.8 ± 15.3
43 (IPR-3147)			Cl	H	2.5 ± 0.3	26.3 ± 6.8

^aRepresentative of at least two independent experiments, where each concentration point is measured in duplicates (mean ± SD).

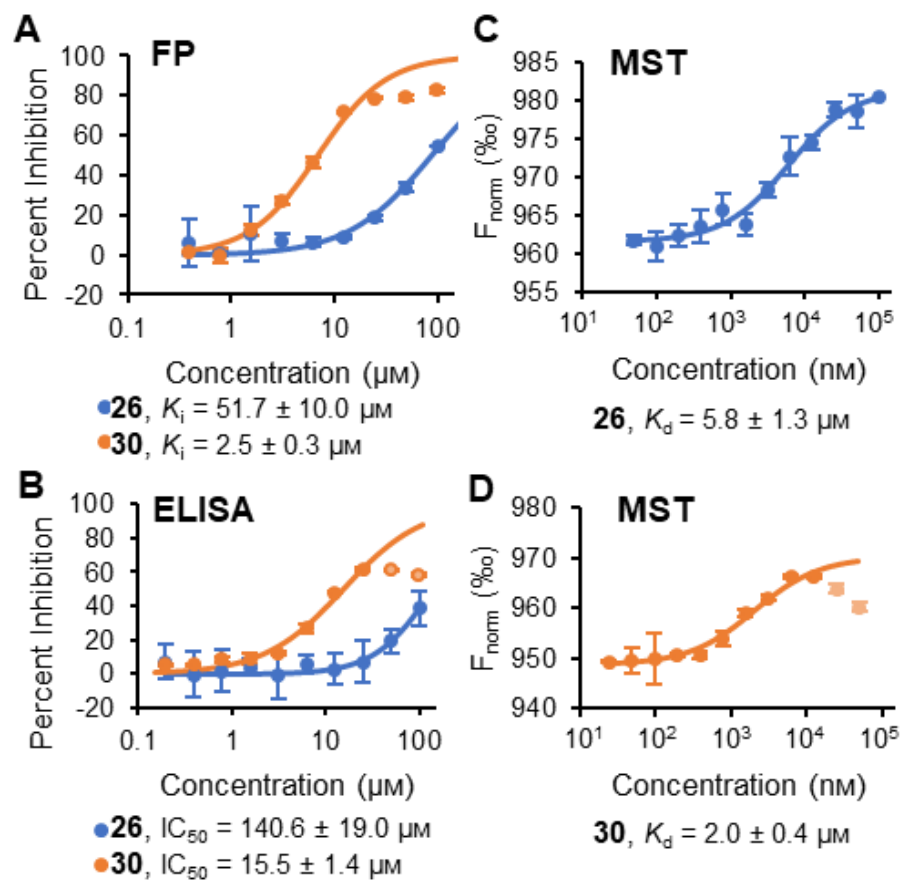


Figure 3.9. Concentration-dependents studies for **26** (IPR-2992) and **30** (IPR-3011). **(A)** Concentration-dependent FP assay measuring the inhibition of uPAR•AE147-FAM peptide interaction by **26** and **30**. **(B)** Concentration-dependent ELISA assay measuring inhibition of uPAR•uPA_{ATF} interaction by **26** and **30**. **(C)** MST experiment was performed with 40 nM NT-495-labeled uPAR and varying concentrations of **26**. **(D)** MST experiment was performed with 40 nM NT-495-labeled uPAR and varying concentrations of **30**.

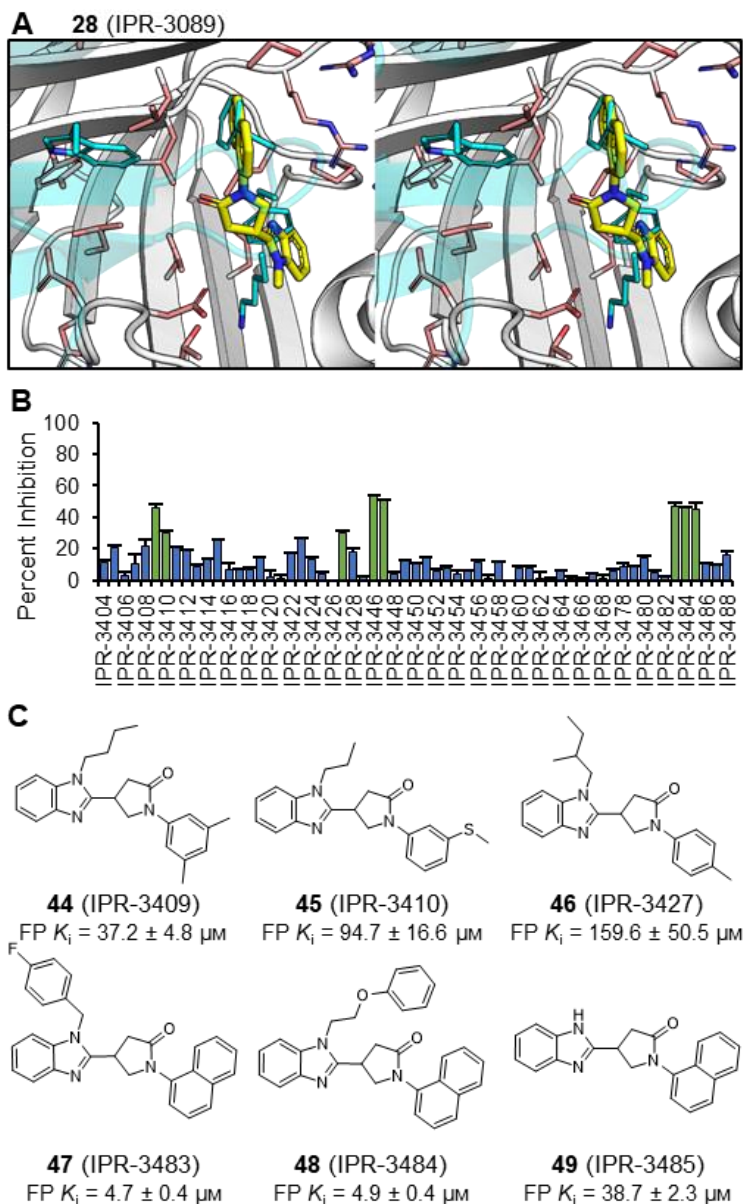


Figure 3.10. Screening the derivatives of compound **28** (IPR-3089). **(A)** The virtual screening binding mode of **28** in the uPAR•uPA binding pocket. The compound is shown in yellow. uPAR is shown in white cartoon, with the side chain of interface residues shown in pink stick. uPA is shown in partial transparent cyan cartoon. The side chain of four interface residues on uPA are shown in stick and colored cyan. **(B)** Derivatives of **28** were screened at a single 50 μM concentration via FP assay in duplicates (mean \pm SD). Further pursued hits are highlighted in green. **(C)** Chemical structures of the pursued derivative hits of **28**.

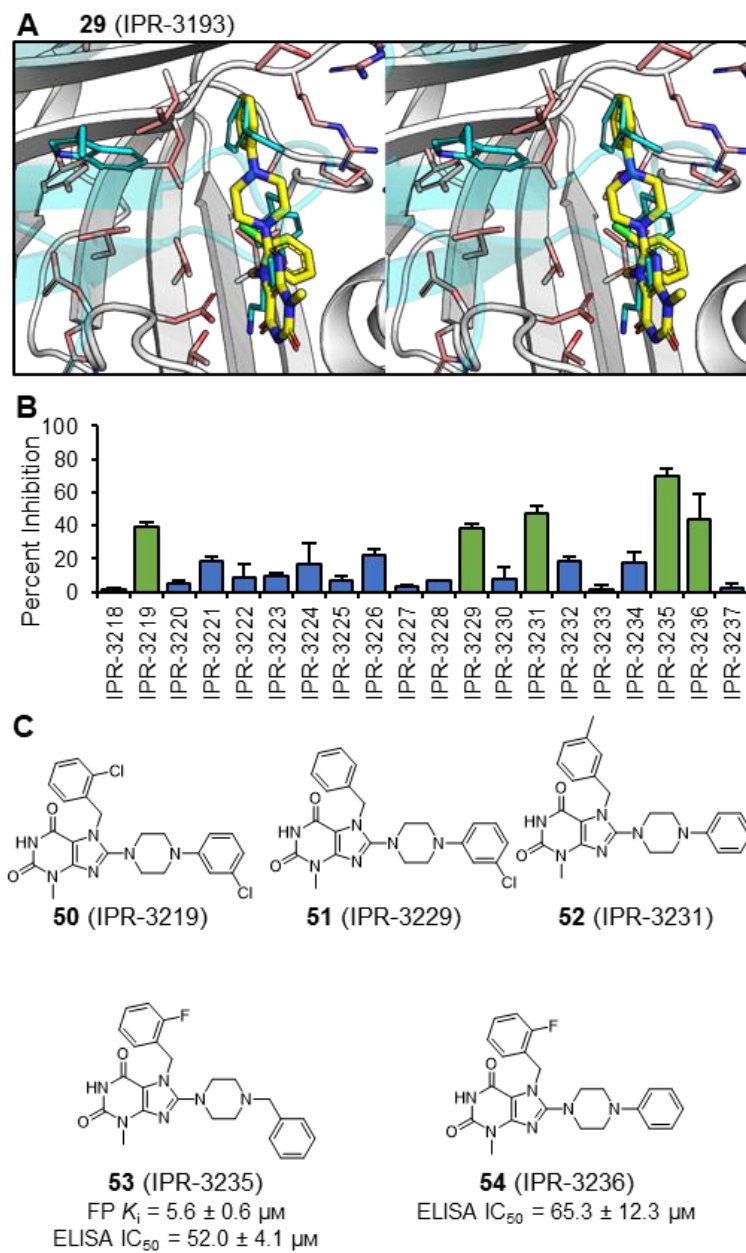


Figure 3.11. Screening the derivatives of compound **29** (IPR-3193). **(A)** The virtual screening binding mode of **29** in the uPAR•uPA binding pocket. The compound is shown in yellow. uPAR is shown in white cartoon, with the side chain of interface residues shown in pink stick. uPA is shown in partial transparent cyan cartoon. The side chain of four interface residues on uPA are shown in stick and colored cyan. **(B)** Derivatives of **29** were screened at a single 50 μM concentration via FP assay in duplicates (mean \pm SD). Further pursued hits are highlighted in green. **(C)** Chemical structures of the pursued derivative hits of **29**.

The top candidates that emerged from this screen were purchased and tested for binding to uPAR. Among them, we found one hit, **1** (IPR-2797), that inhibited a fluorescently-labeled peptide with an inhibition constant K_i of 7 μ M. Analogs of **1** were purchased, confirming the activity of the parent compound and providing an opportunity for structure-activity relationships. Neither **1** nor its derivatives inhibited in our ELISA, which includes the entire protein-protein interaction interface. This suggests that improvements could be made to this method, such as perhaps focusing only on the most critical residues on the receptor. A more stringent threshold for picking interface amino acids could make this possible. Another possibility is that there exists a combination of residues that must be engaged to disrupt a protein-protein interaction, and that compound **1** does not engage the right combination of residues. Regardless, the compound offers an excellent starting point to develop potent small-molecule inhibitors of the tight uPAR•uPA protein-protein interaction. We found the compound to have good solubility, does not react with uPAR, and is not redox active. It is also stable in methanol and buffer as evidenced by LC-MS analysis.

Another strategy that we followed was driven by the hypothesis that small molecules that possess substituents that mimic the position of protein ligand amino acid side chains will more likely disrupt the protein-protein interaction. This is a widely used method, but to the best of our knowledge has never been applied for the discovery of small-molecule inhibitors of tight protein-protein interactions. We resorted to pharmacophore modeling to score compounds based on how effectively they overlap with native protein ligand residue side chains of uPA. This approach resulted in more hit compounds than the screen using receptor amino acids alone. However, close inspection of the structure of these compounds revealed potentially problematic groups, such as a Betti base in compound **6** (IPR-2496) that could result in unstable compounds, or a thiol reactive moiety in compound **10** (IPR-2565) that may lead to adduct formation with nucleophilic residues on uPAR. Among all the compounds, we confirmed that **10** is thiol reactive. Compound **6** was unstable, as expected, despite the single-digit inhibition in both the FP and ELISA assays for a series of derivatives. In fact, we confirmed that the activity of the compound was due to non-specific reactivity through the synthesis of **18** (IPR-2804), a derivative that lacked the Betti base. Despite the lack of obvious unstable or reactive moieties for compound **5** (IPR-2477), we found it to be unstable and its activity is likely due to assay interference or reactivity with assay or protein. Compounds **8** (IPR-2529) and **9** (IPR-2532) were the most robust compounds we identified. Only compound **9** inhibited in both the FP and ELISA suggesting that it could be a good starting point for the development of uPAR•uPA inhibitors. Its large size, however, may make it difficult to optimize. Compound **8** did not inhibit uPAR•uPA in our ELISA suggesting that the compound

binds, but it may not effectively mimic uPA residues. It is also possible that the compound binds to residues on uPAR that negate the benefits of mimicking uPA.

Finally, we wondered whether the use of interface residues on both uPAR and uPA could lead to better inhibitors from virtual screening. We combined our fingerprint and pharmacophore methods to rank-order compounds docked to uPAR. It is interesting that this method led to even more hit compounds than using fingerprint or pharmacophore alone. A total of seven hits were identified. Despite the initial concern that three of the compounds had potentially problematic α,β -unsaturated carbonyls, such as in **23** (IPR-2986), **24** (IPR-2987), or **27** (IPR-2993), none of the compounds were found to be thiol reactive. One compound had an acylhydrazine moiety that could also be unstable at low pH, although our work is done at pH 7 suggesting that the compound should be stable. An interesting feature of these compounds compared to those that emerged from using strictly the pharmacophore method is that they had fewer rotatable bonds overall, and two compounds, namely **26** (IPR-2992) and **28** (IPR-3089) were fragment-like. Compound **26** was particularly interesting as it inhibited, albeit weakly, in both our FP and ELISA assays. Starting with **26**, we followed an analog-by-catalog approach and purchased several derivatives. Among the derivatives, we discovered several compounds, including **30** (IPR-3011), which exhibited substantially higher binding affinity than the parent fragment-like compound. We confirmed direct binding of both **26** and **30** using microscale thermophoresis with K_d values that were similar to the K_i s values measured by FP. **30** also possessed substantially better IC_{50} s (single-digit micromolar range) in the disruption of the full uPAR•uPA interaction. Future optimization will focus on improving solubility of these derivative compounds and exploring additional substituents for **26**. The methyl group located on the quinoline ring points towards the side chain of a critical tryptophan on uPA. The introduction of moieties that mimic the tryptophan side chain may result in substantially greater potency.

In sum, we present a new approach to identify small-molecule inhibitors of tight protein-protein interactions that uses the native ligand's pairwise intermolecular interactions with the receptor of a protein-protein interaction. When combined with a pharmacophore approach that uses the native protein ligand interface amino acids, we identified robust small-molecule inhibitors of the tight uPAR•uPA. To the best of our knowledge, this is the first example of a virtual screen that uses the crystal structure of a tight protein-protein interaction and identified single-digit micromolar small-molecule inhibitors. These results suggest that while commercial libraries do not cover chemical space that is typical of protein-protein interaction inhibitors, it is possible to identify robust starting points that could be used to develop small-molecule inhibitors of tight protein-protein interactions. The results also show that virtual screening is also prone to nuisance

compounds as several of the small molecules that initially showed promising activity were working through a non-specific mechanism. Finally, it is worth mentioning that small-molecule inhibitors that emerged from this work are structurally distinct from inhibitors that we previously identified for uPAR. We compared the structure of compounds **1-54** to our previously reported uPAR•uPA inhibitors IPR-803 [165] and IPR-1110 [167, 168]. The similarity between these compounds was assessed using atom triplet Daylight fingerprints. We find that generally, the Tanimoto similarity between compounds **1-54** and our previously described inhibitors range from 0.05 to 0.10.

3.4 MATERIALS AND METHODS

3.4.1 Virtual Screening. A set of commercially-available compounds from ChemDiv Inc. (San Diego, CA), ChemBridge Corporation (San Diego, CA), Life Chemicals (Munich, Germany), Princeton BioMolecular Research Inc. (Princeton, NJ), Specs (Zoetermeer, Netherlands), and Vitas-M Laboratory Ltd (Hong Kong) were retrieved from ZINC [169]. Small molecules in this library possessing pan-assay interference compound (PAINS) [170] or rapid elimination of swill (REOS) [171] moieties were filtered out using the Canvas package in Schrödinger (Schrödinger LLC, New York, NY, 2015). This resulted in a compound library of approximately 5.1 million small molecules. Individual MOL2 formatted files were converted to PDBQT format using the *prepare_ligand4.py* script in MGLTools [172].

The structure of the uPAR•uPA complex (PDB ID: 3BT1) was retrieved and prepared using Protein Preparation Wizard in Schrödinger [173]. Bond orders were assigned, hydrogen atoms were added, and disulfide bonds were created. Vitronectin (chain B) was removed and the missing loop at residues Arg-83 and Ala-84 were introduced using the Prime module. The resulting structure was protonated at pH 7.0 using PROPKA [124] and separated into its respective monomeric chains. The uPAR structure (chain U) was converted to PDBQT format using the *prepare_receptor4.py* script in MGLTools.

The compound library was docked to the prepared uPAR structure using AutoDock Vina [172]. The binding pocket was centered at the uPAR•uPA interface with a box with dimensions of 21 Å × 21 Å × 21 Å. All other parameters were set to default values. The docked conformations were converted back to MOL2 format using in-house Python scripts for additional analysis.

3.4.2 uPAR Interface Residues. To find compounds that overlapped with residues on uPAR in the uPAR•uPA complex, we resorted to a fingerprint approach that utilizes interaction energies between the receptor and ligand. We determined the interaction energies of each docked compound to individual residues of uPAR using the Generalized Born Surface Area (GBSA) method in the Amber14 and AmberTools15 software packages [126]. Each docked compound was

assigned Gasteiger charges and gaff [128] atom types using the *antechamber* program [129]. Additional force field parameters were generated using the *parmchk* program. Topology and coordinate files for the docked complex and individual receptor and ligand were generated with ff14SB [131] and gaff [128] force fields using the *tleap* program. These topology and coordinate files were used as inputs to calculate the free energies and per-residue decomposition energies in the *MMPBSA.py* script [136]. The *MMPBSA.py* script was modified to include the missing atom radius for iodine atoms [174]. The calculation using the Generalized Born (GB) method was performed with *sander* and Onufriev's GB model [137, 138]. Solvent-accessible surface area (SASA) calculations were switched to the icosahedron (ICOSA) method, where surface areas are computed by recursively approximating a sphere around an atom, starting from an icosahedron. Salt concentration was set to 0.1 M. Compounds with combined internal and solvation terms (ΔE_{GBTOT}) greater than $-5.0 \text{ kcal}\cdot\text{mol}^{-1}$ were discarded.

For each docked compound, we generate a one-dimensional array with length equal to the total number of residues of the uPAR structure. In this vector, each position corresponds to an individual residue of uPAR. Each position is assigned a value of '1' (ON) or '0' (OFF) based on the residue decomposition energy at that position and acts as a fingerprint for that compound. If the energy at the given residue is less than $-1.0 \text{ kcal}\cdot\text{mol}^{-1}$, we assign the position a value of '1'. Otherwise, we assign the position a value of '0'. Residues for the fingerprints were identified from two sources: (i) an experimentally-determined alanine scanning of uPAR from Gårdsvoll and coworkers [164]; and (ii) a previously described molecular dynamics (MD) simulation of the uPAR•uPA complex [168]. Similar to the construction of the compound-specific bitwise arrays, we create vectors for each type of fingerprint where each position corresponds to an interaction energy of the uPAR•uPA complex. In the vector corresponding to the experimental alanine scan, a position was assigned a value of '1' if the $\Delta\Delta G$ at that residue is greater than $1.0 \text{ kcal}\cdot\text{mol}^{-1}$ and '0' otherwise. In the vector corresponding to the per-residue decomposition energies, a position is assigned a value of '1' if the total energy (ΔE_{GBTOT}) at that residue is less than $-1.0 \text{ kcal}\cdot\text{mol}^{-1}$ and '0' otherwise.

In both fingerprints, only a small portion of uPAR will have values of '1' with its native ligand uPA. Therefore, we reduce the length of each fingerprint to only include positions with '1' bits in the uPAR•uPA complex. For each docked compound, we calculate the Tanimoto distance between the fingerprints of the complex and the compound in a bitwise manner. The fingerprint of the uPAR•uPA complex consists of only '1' bits. Thus, this distance can be simply calculated by summing the number of '1' bits in the compound fingerprint and dividing by the length of the

fingerprint. Compounds were rank-ordered based on their Tanimoto distance, and in cases where compounds had the same Tanimoto distance, we used ΔE_{GBTOT} to rank these compounds.

3.4.3 uPA Interface Residues. A pharmacophore-based approach was used to identify docked compounds that overlapped with and mimicked interface residues on uPA. We used four residues of uPA at the uPAR•uPA interface: Lys-23, Tyr-24, Phe-25, and Trp-30. For each residue, we defined a pharmacophore hypothesis corresponding to the physiochemical properties of the individual residue's side chain using the Phase package in Schrödinger [115, 116]. Phase has six built-in types of pharmacophore features: (i) hydrogen bond acceptor, (ii) hydrogen bond donor, (iii) hydrophobe, (iv) negative ionizable, (v) positive ionizable, and (vi) aromatic ring. We assigned a positive charged feature to the ϵ -amine on Lys-23 and aromatic rings features to the aromatic rings of Tyr-24, Phe-25, and Trp-30. A single pharmacophore feature was assigned to the benzene rings of Tyr-24 and Phe-25, while two separate pharmacophores were assigned to the pyrrole and benzene rings of the bicyclic indole on Trp-30. We searched for compounds containing ligand moieties that matched a corresponding pharmacophore feature. A compound that matched either of the two aromatic pharmacophore features on Trp-30 was considered to overlap and mimic the residue. Each pharmacophore feature was screened independently of one another. The aromatic and positively charged pharmacophores are represented as spheres centered on the side chain moiety with radii of 1.5 Å and 0.75 Å, respectively. Compound conformers from virtual screening were used to identify matches without refinement using Phase's default fitness function. In this scoring function, three factors are used to describe the degree to which a compound matches a pharmacophore feature: (i) the site score, which describes how well the compound superimposes the pharmacophore feature, (ii) the vector score, which describes the cosine angle between the normal vector of an aromatic ring on the compound with an aromatic feature, and (iii) the volume score, which describes the proportion of the total volume of the pharmacophore feature overlapped by the compound. Compounds with either root-mean-square deviation (RMSD) overlap greater than 1.2 Å or did not overlap with the pharmacophore feature were discarded. For the aromatic pharmacophores, no consideration was given to the angle between the normal vectors of an aromatic feature and the orientation of an aromatic ring. All other parameters were set at default values. The remaining compounds that matched a given pharmacophore was retained without sorting compounds by Phase's internal scoring function. Compounds that matched 3 of the 4 residues were retained.

3.4.4 Selection of Compounds. The top-ranking compounds following virtual screening using uPAR and uPA residues were retrieved and clustered using the Canvas package in Schrödinger. A hashed binary fingerprint corresponding to atom triplets of Daylight invariant atom

types were generated for these top-ranking compounds. Compounds were then hierarchically clustered using their atom triplet fingerprints and average linkage clustering. The Tanimoto similarity between a pair of fingerprints was used as the distance metric. Compounds corresponding to the cluster centers from hierarchical clustering were purchased for experimental validation.

3.4.5 Fluorescence Polarization (FP) Assay. Polarized fluorescence intensities were measured using EnVision Multilabel plate readers (PerkinElmer, Waltham, MA) with excitation and emission wavelengths of 485 and 535 nm, respectively [165]. Samples were prepared in Thermo Scientific Nunc 384-well black microplate in duplicates. First, the compounds were serially diluted in DMSO and further diluted in 1× PBS buffer with 0.01% Triton X-100 for a final concentration of 200 – 0.2 μM . Triton X-100 was added to the buffer to avoid compound aggregation. 5 μL of the compound solution and 40 μL of PBS with 0.01% Triton X-100 containing uPAR was added to the wells and incubated for at least 15 min to allow the compound to bind to the protein. Finally, 5 μL of fluorescent AE147-FAM peptide solution was added for a total volume of 50 μL in each well resulting in final uPAR and peptide concentrations of 320 nM and 100 nM respectively. The final DMSO concentration was 2% v/v, which had no effect on the binding of the peptide. Controls included wells containing only the peptide and wells containing both protein and peptide each in duplicates to ensure the validity of the assay. A unit of millipolarization (mP) was used for calculating percentage inhibition of the compounds. When compounds were insoluble and visible precipitation was observed, the data points at the high concentrations were not included in the calculation of IC_{50} values. Inhibition constants were calculated from the IC_{50} values using the K_i calculator available at http://sw16.im.med.umich.edu/software/calc_ki/.

3.4.6 Microtiter-Based ELISA for uPAR•uPA. uPAR without the glycosylphosphatidylinositol (GPI) anchor was obtained by a purification process as previously described [175]. High-binding microplates (Greiner Bio-One, Kremsmünster, Austria) were incubated for 2 h at 4 °C with 100 μL of 4 $\mu\text{g}\cdot\text{mL}^{-1}$ of uPA_{ATF} in PBS for immobilization as previously described [165]. The plate was washed with 0.05% Tween-20 in PBS buffer between each step. A 1:1 mixture of Superblock buffer in PBS (Thermo Fisher Scientific, Inc. Waltham, MA) with 0.04 M NaH_2PO_4 and 0.3 M NaCl buffer was used for blocking at room temperature for 45 min. Following removal of the blocking buffer and washing, 100 μL of 0.85 nM uPAR in PBS with 0.01% triton X-100 was added with 100 to 0.4 μM compounds in 1% v/v DMSO. Following incubation for 30 min and subsequent washing steps, biotinylated human uPAR antibody (1:3000 dilution of 0.2 $\text{mg}\cdot\text{mL}^{-1}$ BAF807, R&D Systems, Minneapolis, MN) in PBS containing 1% bovine serum albumin (BSA) was added to the wells (100 μL /well) and incubated for 1 h to allow for the detection of bound uPAR. Following washing, streptavidin-horseradish-peroxidase in PBS

containing 1% BSA was added to the wells and incubated for 20 min. The signal developed in the presence of 3,3',5,5'-tetramethylbenzidine (TMB) in phosphate-citrate buffer (pH 5) and hydrogen peroxide was stopped by adding H₂SO₄ solution and was detected using a SpectraMax M5e (Molecular Devices, Sunnyvale, CA). When compounds were insoluble and visible precipitation was observed, the data points at the high concentrations were not included in the calculation of IC₅₀ values.

3.4.7 Microscale Thermophoresis (MST). uPAR was labeled with NT-495 fluorescent dye (Nanotemper, Munich, Germany) according to the manufacturer's instructions. Compounds were serially diluted in DMSO and further diluted in PBS buffer with 0.025% v/v Tween-20. 10 μL of fluorescently-labeled uPAR and 10 μL of compound solution were combined to final concentrations of 40 nM fluorescently-labeled uPAR, 2% v/v DMSO and compound concentrations ranging from 200 μM to 0.1 μM. The protein-compound solution was incubated for 10 min at room temperature in the dark, before being taken up in Monolith NT.115 series standard-treated capillaries. The capillaries were measured on Monolith NT.115 (Nanotemper, Munich, Germany) at 25 °C, with LED power at 40% and MST power at 40%, and the MST was measured for 30 s. The data was analyzed using the "Thermophoresis with T-jump" function within the NanoTemper Affinity Analysis version 2.0.2 software (Nanotemper, Munich, Germany). The data was then fit with the "K_d Model" function within the software to calculate the K_d.

3.4.8 (E)-2-(4-mercaptostyryl)-1,3,3-trimethyl-3H-indol-1-ium (MSTI) Assay. MSTI assay was performed according to the manufacturer's recommendations (Kerafast, Inc. Boston, MA) [166]. A 10 mM solution of acetyl-MSTI was added to 50 mM PBS buffered solution at pH 12.0 with 50% v/v methanol in a ratio of 1:10 v/v. After mixing and incubating for 2 min at room temperature, the solution was diluted with 50 mM PBS at pH 7.4, containing 0.01% NP-40 and 5% v/v methanol to generate a final concentration of 30 μM MSTI at pH 7.4. 19.6 μL aliquots of the MSTI solution was dispensed in 384-well flat-bottom black plate and 0.4 μL of 5 mM compounds in DMSO were added make a 100 μM final concentration. Unreacted MSTI solution without added compound, but with equal amount of DMSO, was used as a negative control, while inactivated acetyl-MSTI solution with DMSO was used as a positive control. The plate was then incubated with shaking for 30 min at room temperature and the fluorescence intensities were measured using a Flexstation 3 plate reader (Molecular Devices, Sunnyvale, CA) at excitation and emission wavelengths of 510 and 650 nm, respectively.

3.4.9 Horseradish Peroxidase-Phenol Red (HRP-PR) Redox Activity Assay. HRP-PR assay was performed according to the published protocol [176, 177]. In brief, 20 μL of 300 μM compounds in 3% v/v DMSO in Hank's balanced salt solution (HBSS) (Cat. No. SH30268.02;

HyClone, Logan, UT) was dispensed into a 384-well clear, flat-bottomed polystyrene plate (Cat. No. 781101; Greiner Bio-One, Monroe, NC). Controls with 3% v/v DMSO and 300 μM H_2O_2 were dispensed. 20 μL of 2.4 mM fresh dithiothreitol (DTT) in HBSS was added to each well. For the H_2O_2 controls, 20 μL HBSS with no DTT was added. After 5 min incubation at RT, 20 μL solution of 300 $\mu\text{g}\cdot\text{mL}^{-1}$ Phenol Red (Cat. No. P-2417; Sigma-Aldrich, St. Louis, MO), 180 $\mu\text{g}\cdot\text{mL}^{-1}$ HRP (Cat. No. P-2088; Sigma-Aldrich, St. Louis, MO) was added. After 20 min incubation at room temperature, the absorbance was read at 610 nm on a SpectraMax M5e (Molecular Devices, Sunnyvale, CA).

3.4.10 High-Performance Liquid Chromatography-Mass Spectrometry (HPLC-MS).

Compounds at 200 μM were incubated in methanol, PBS, or 30 μM uPAR in PBS for 1 h at room temperature. The samples were injected onto a Kinetex 2.6 μm XB-C18 100 Å column (Cat. No. 00B-4496-E0; Phenomenox, Torrance, CA) on an Agilent 6130 Quadrupole LC/MS system (Agilent, Santa Clara, CA). Compounds **6** (IPR-2496), **12** (IPR-2605), and **29** (IPR-3193) were eluted by a linear gradient from Buffer A (H_2O) to Buffer B (acetonitrile, 5 mM NH_4OAc) over 15 min. Compounds **1** (IPR-2797), **8** (IPR-2529), **9** (IPR-2532), **26** (IPR-2992), and **28** (IPR-3089) were eluted by a linear gradient from Buffer A (H_2O , 0.1% formic acid) to Buffer B (acetonitrile, 0.1% formic acid) over 15 min. Column elution was tracked by UV absorption at 256 nm and the masses were detected by positive ion mode.

Chapter 4

CHEMICAL SPACE OVERLAP WITH CRITICAL PROTEIN-PROTEIN INTERFACE RESIDUES IN COMMERCIAL AND SPECIALIZED SMALL-MOLECULE LIBRARIES

4.1 INTRODUCTION

There is growing interest in applying computational methods for the discovery of small-molecule PPI inhibitors, particularly for tight secondary and tertiary interactions. We have successfully used virtual screening to identify inhibitors of the tight tertiary interaction between the urokinase receptor uPAR and its protein ligand uPA with compounds ranging from sub-micromolar to micromolar affinity [165, 168, 178]. In one case, we screened multiple structures that were sampled from explicit-solvent molecular-dynamics simulations and identified a sub-micromolar affinity compound [165]. The predicted structure of the compound was independently confirmed by X-ray crystallography [179]. In the previous chapter, we introduced a fingerprint method that uses the native protein ligand as a guide to identify small molecules that mimic the interaction of the protein ligand [178]. Using the fingerprint method, we identified several hit compounds with single-digit micromolar affinities. Finding highly potent inhibitors of tight PPIs by virtual screening of commercial libraries remains extremely challenging, but we have shown that quality hit compounds with single-digit micromolar binding affinities can be identified.

Although progress has been made in the development of scoring and docking methods to enrich compound collections for the discovery of PPI inhibitor hit compounds, relatively little work has been done on the suitability of existing commercial and specialized collections for PPI drug discovery. Databases such as iPPI [180], 2P2I [181], and TIMBAL [182] have been created to explore additional chemotypes for PPIs. A few studies have explored whether commercial and specialized libraries contain small molecules that are suitable for disrupting PPIs [183-185]. These studies suggest that small-molecule PPI inhibitors tend to cover different chemical space than enzyme inhibitors. Another approach to determine whether a compound collection is enriched with compounds that are suitable for disrupting PPIs is to explore how effectively compounds mimic the sidechains of critical amino acids of the PPI ligand protein, or whether compounds engage critical residues on the PPI receptor protein [186]. Small molecules that better mimic or engage critical residues are expected to be more effective inhibitors of PPIs [30-32].

In this chapter, we explored how effectively small molecules in a commercial library (ChemDiv), a collection of diversity-oriented synthesis (DOS) libraries, and the “Screenable Chemical Universe Based on Intuitive Data Organization” (SCUBIDOO) compound collection mimic the positions of critical residues at tight protein-protein interfaces. We employed a combined

docking and pharmacophore approach to measure overlap between chemical structures and the sidechains of interface residues. We selected three tight PPIs for which extensive alanine-scanning studies have identified critical hot spot residues; these include the interaction between the urokinase receptor (uPAR) and its serine proteinase ligand urokinase (uPA), the transcription factor TEAD and its co-activator Yap (TEAD•Yap), and the α - and β -subunits of the voltage-gated calcium channels ($\text{Ca}_v\alpha\text{•Ca}_v\beta$). Our findings suggest that smaller conformationally-restricted compounds show excellent overlap with hot spots. In a proof-of-concept study, we experimentally validated the compounds identified from virtual screening of a library of conformationally-restricted commercially-available compounds against the PPIs of uPAR•uPA and TEAD•Yap.

4.2 RESULTS

4.2.1 Analysis of Compound Collection Physicochemical Properties. Several types of chemical library have been created over the years that may be suitable in identifying PPI inhibitors. Among them, diversity-oriented synthesis (DOS) represents a strategy to efficiently generate compound collections with a high degree of structural diversity [187, 188] and produce new biological probes [42, 189-194]. The DOS approach aims to achieve building-block, functional group, stereochemical, and skeletal diversity, with DOS compounds occupying a middle ground between the structural complexity of natural products and the efficiency of commercially-available synthetic libraries [195]. In addition to commercially available and DOS compounds, there is an increasing number of specialized libraries that have been constructed for use in virtual screening. These libraries were designed to overcome some of the shortcomings of commercial libraries, namely the limited chemical space commercial libraries cover and the large number of nuisance compounds [196-198]. The Screenable Chemical Universe Based on Intuitive Data Organization (SCUBIDOO) library [199] is one example. It consists of compounds constructed by combining building blocks from 58 organic reactions known to the pharmaceutical field.

To explore the structural overlap between small molecules and amino-acid sidechains at protein-protein interfaces, three compound collections from both commercial and non-commercial sources were selected. The first was the commercially-available library from ChemDiv, Inc. (ChemDiv), which is frequently used in compound screening. The second was a collection of DOS libraries [194, 200] from the Broad Institute (DOS), and the third was SCUBIDOO [199]. The physicochemical properties of the compounds in each of the three collections were determined (**Fig. 4.1**). Compounds in DOS had a mean molecular weight (MW) of 512 ± 94 Da, which is larger than the mean molecular weight of compounds in ChemDiv (MW = 410 ± 74 Da) and SCUBIDOO (MW = 327 ± 49 Da).

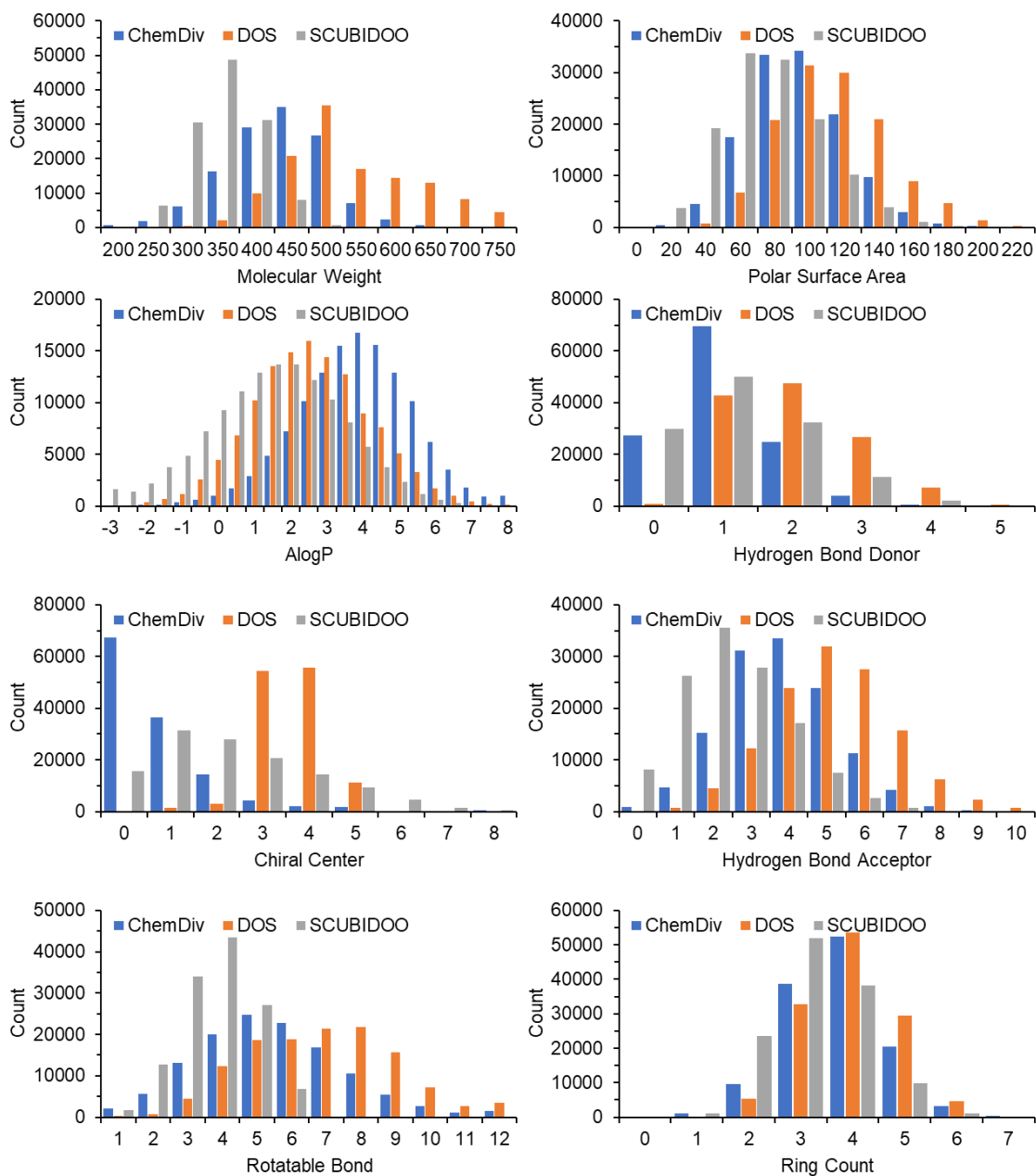


Figure 4.1. Histograms of individual physicochemical properties of the ChemDiv, DOS, and SCUBIDOO compound collections. Distributions highlight differences in size, flexibility, and complexity.

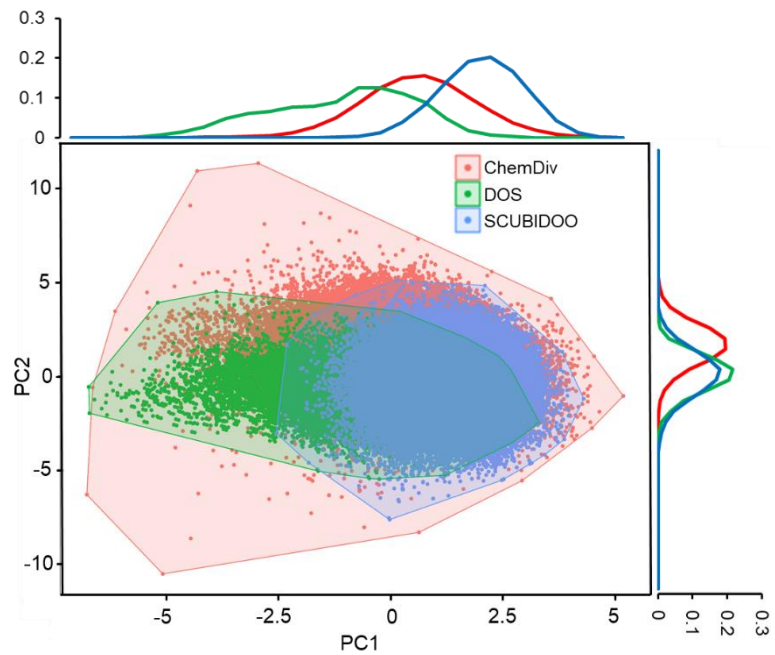


Figure 4.2. Principal component analysis (PCA) of physicochemical properties. Compounds from each of the three collections were projected onto the first two principal components, each of which is also represented by its marginal distribution on a per-collection basis.

Table 4.1. Parameters for principal component analysis for each of the eight input descriptors across the first three principal components.

	Scale	PC1	PC2	PC3
AlogP	1.907	-0.157	0.692	-0.056
Hydrogen-bond acceptor	1.860	-0.439	-0.082	-0.241
Hydrogen-bond donor	0.965	-0.300	-0.338	-0.061
Molecular weight	106.326	-0.493	0.135	0.189
Polar surface area	32.87	-0.448	-0.152	-0.330
Rotatable bond	2.220	-0.421	-0.044	-0.061
Chiral center	1.719	-0.132	-0.407	0.733
Ring count	0.975	-0.233	0.438	0.499
Standard deviation		1.865	1.210	1.064
Component variance		0.435	0.183	0.142
Cumulative variance		0.435	0.618	0.760

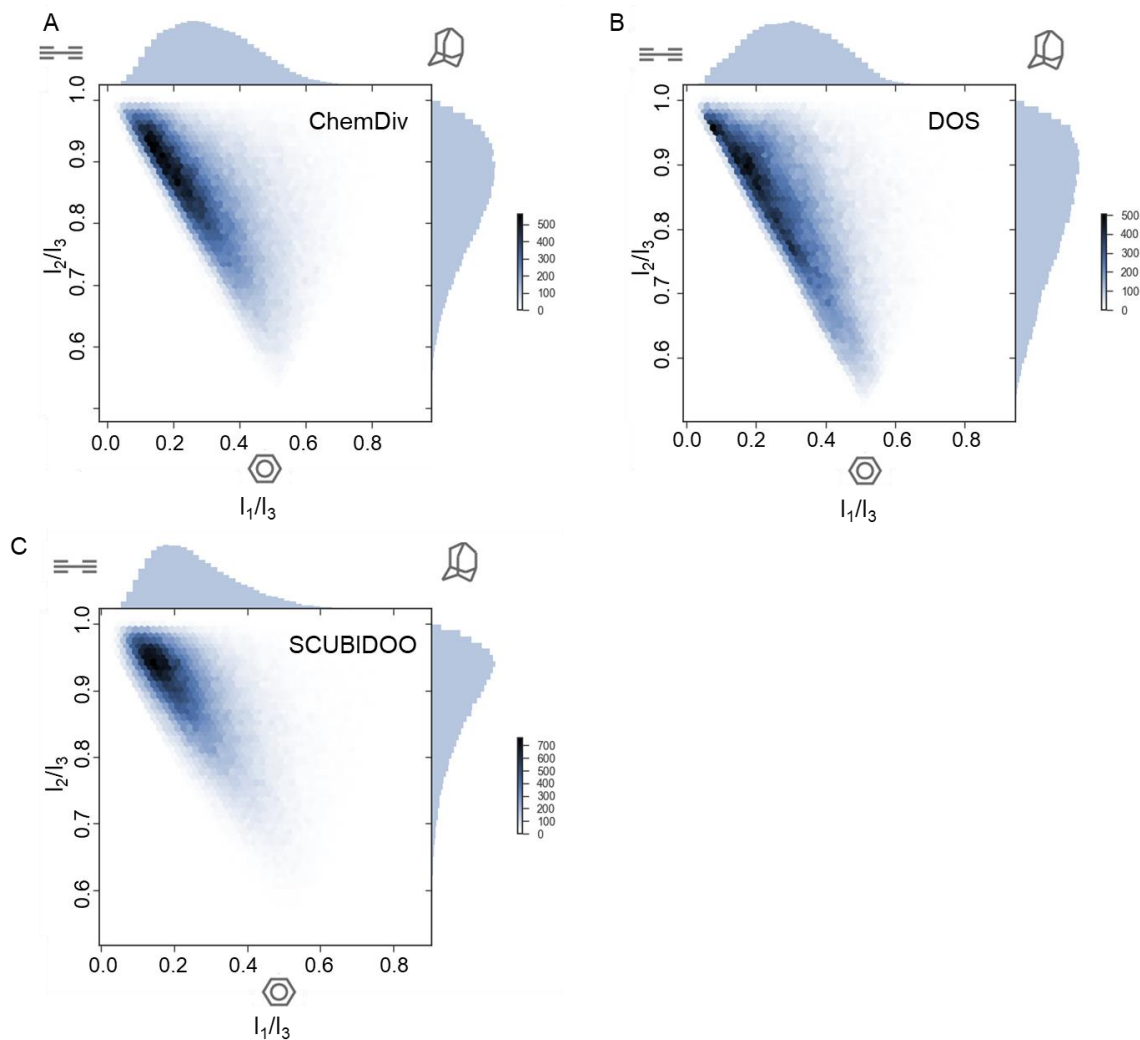


Figure 4.3. Principal moments of inertia (PMI) illustrate the shape diversity of three compound collections. (A) ChemDiv, (B) DOS, (C) SCUBIDOO. The top left-hand corner represents a linear molecule (e.g., diacetylene), the top right-hand corner represents a spherical molecule (e.g. adamantane) and the bottom corner represents a disc-like molecule (e.g., benzene).

Lipophilicity, characterized here by computed AlogP values, plays a crucial role in determining solubility. Compounds in ChemDiv had a mean AlogP of 3.6 ± 1.5 . ChemDiv compounds are predicted to be generally less soluble than compounds from DOS or SCUBIDOO considering their substantially lower mean AlogP values of 2.3 ± 1.6 , and 1.3 ± 1.8 , respectively. To gain insight into the three-dimensional characteristics and complexity of compounds in these collections, the number of chiral centers was counted. DOS compounds had generally more chiral centers (3.6 ± 0.8), compared with SCUBIDOO (2.4 ± 0.7) and ChemDiv compounds (0.7 ± 1.2). These data confirmed that DOS compounds are generally more stereochemically complex than SCUBIDOO compounds. ChemDiv compounds showed a remarkably low number of chiral centers per compound compared with both DOS and SCUBIDOO.

A measure of flexibility is the number of rotatable bonds in a compound. There were more rotatable bonds on average per compound in DOS (RB = 6.8 ± 2.1) compared with SCUBIDOO (RB = 3.8 ± 1.1) and ChemDiv (RB = 5.5 ± 2.1). The molecular weight of a compound was generally correlated with the number of rotatable bonds in the DOS (Pearson's $r = 0.68$) and ChemDiv collections ($r = 0.55$), but not in the SCUBIDOO library ($r = 0.13$). There was a similar trend in the correlations between molecular weight and number of chiral centers in DOS ($r = 0.31$), ChemDiv ($r = 0.20$), and SCUBIDOO ($r = 0.08$). Similarly, the number of chiral centers and rotatable bonds was only correlated in the DOS collection ($r = 0.27$).

Polar surface area (PSA) is a commonly-used descriptor to provide insight into compound permeability and oral bioavailability [201]. In addition, PSA is often used in combination with the number of rotatable bonds to reflect molecular flexibility [202]. Compounds in DOS had the highest PSA values with a mean of $104 \pm 31 \text{ \AA}^2$, followed by ChemDiv (PSA = $84 \pm 27 \text{ \AA}^2$) and SCUBIDOO (PSA = $66 \pm 28 \text{ \AA}^2$). We found that PSA and the number of rotatable bonds were most strongly correlated in DOS ($r = 0.60$) and ChemDiv ($r = 0.40$), but not in SCUBIDOO ($r = -0.02$).

The number of hydrogen-bond donors (HBDs) and hydrogen-bond acceptors (HBAs) are important parameters related to compound polarity and membrane permeability. It has been suggested that the number of HBDs may be more important than the number of HBAs to enhance bioavailability and membrane permeability of lead compounds [201, 203, 204]. We found that DOS compounds had the highest number of HBDs and HBAs (HBD = 2.0 ± 0.9 ; HBA = 5.2 ± 1.6), followed by ChemDiv (HBD = 1.0 ± 0.8 ; HBA = 3.8 ± 1.4) and SCUBIDOO (HBD = 1.2 ± 0.9 ; HBA = 2.6 ± 1.4). Among all the physicochemical properties that we have considered, only the ring count was similar among the three collections. Since SCUBIDOO was built by limiting the number of building blocks, compounds in SCUBIDOO have lower molecular weight, rotatable bonds, and chiral centers compared to compounds in DOS and ChemDiv.

We next compared the distributions of these eight physicochemical properties of the three collections using principal component analysis (PCA). Compounds from each of the three collections were projected onto the first two principal components (**Fig. 4.2**). In the first principal component, the distributions were separated such that there are distinct peaks for each of the three sources ($PC1_{DOS} = -1.6 \pm 1.7$; $PC1_{ChemDiv} = 0.1 \pm 1.3$; $PC1_{SCUBIDOO} = 1.5 \pm 0.9$). This effect is seen qualitatively in the marginal distributions for PC1 for each compound collection, where each set was projected onto different ranges of the first principal component. In the second principal component, however, the marginal distributions of DOS ($PC2_{DOS} = -0.5 \pm 0.9$) and SCUBIDOO ($PC2_{SCUBIDOO} = -0.5 \pm 1.1$) overlapped and each partially overlapped with the marginal distribution of ChemDiv ($PC2_{ChemDiv} = 0.9 \pm 1.0$). The loadings of the PCA are provided to illustrate the relative contributions of each of the eight input descriptors (**Table 4.1**). The total variance explained by the first two principal components were 44% and 18%, respectively. When the third principal component is included, the cumulative variance explained reached 76%.

To gain further insight into the structure of compounds in these three collections, the molecular shape diversity of each was evaluated using principal moments of inertia (PMI) (**Fig. 4.3**). PMI plots represent the shape distribution of a collection of molecules. The three vertices of the triangular plot represent the extremes of molecular geometry. The top-left, top-right, and bottom corners correspond to small molecules with linear (e.g., diacetylene), spherical (e.g., adamantane), and disc-like (e.g., benzene) shapes, respectively. Compounds in SCUBIDOO ($I_1/I_3 = 0.26 \pm 0.12$; $I_2/I_3 = 0.88 \pm 0.08$) were predominantly linear compared to compounds in ChemDiv and DOS. ChemDiv ($I_1/I_3 = 0.31 \pm 0.13$; $I_2/I_3 = 0.84 \pm 0.09$) and DOS ($I_1/I_3 = 0.33 \pm 0.12$; $I_2/I_3 = 0.82 \pm 0.09$) compounds were primarily along the diagonal between linear and disc-like structures. Despite the substantial difference in the number of chiral centers and rotatable between DOS (chiral center = 3.6 ± 0.8 ; RB = 6.8 ± 2.1) and ChemDiv (chiral center = 0.7 ± 1.2 ; RB = 5.5 ± 2.1), there was little difference in the shape diversity between the two collections.

4.2.2 Compound Overlap with Protein-Ligand Hot Spots at Protein-Protein Interaction Interfaces. We wondered if the chemical scaffolds found in each compound collection would yield compounds that could overlap and mimic the physicochemical properties of amino acid sidechains at protein-protein interfaces. To explore this question, we selected three high-affinity ($K_d < 100$ nM) PPIs with three distinct binding motifs: (i) a β -turn motif at the interface between the urokinase receptor and its serine proteinase ligand urokinase (uPAR•uPA); (ii) a twisted-coil motif at the Ω -loop between the transcription factor TEAD and its co-activator Yap (TEAD•Yap); and (iii) an α -helix motif between the α - and β -subunits of the calcium channel (Cav α •Cav β).

We first docked compounds from the ChemDiv, DOS, and SCUBIDOO to the protein receptor at each of the three interaction interfaces. A ligand-based pharmacophore approach was used to identify compounds that overlap with interface residues on the protein ligand. Six possible pharmacophore features can be assigned to the sidechains of protein residues: hydrogen bond acceptor (A) and donor (D), hydrophobic group (H), negatively (N) and positively (P) charged group, and aromatic ring (R). These pharmacophore sites are characterized by type, location, and, if applicable, directionality.

4.2.3 uPAR•uPA. The uPAR•uPA interface consists primarily of a β -turn on the protein ligand uPA ensconced in a large pocket on the protein receptor uPAR, leading to an interaction that is both tight ($K_d = 1$ nM) and stable ($k_{off} = 10^{-4}$ s⁻¹) (**Fig. 4.4A** and **4.4B**). At the uPAR•uPA interface, the sidechain of five hot-spot residues from uPA extend into the hydrophobic pocket of uPAR: Lys-23, Tyr-24, Phe-25, Ile-28, and Trp-30 (**Fig. 4.4C**) [157]. We used pharmacophores to represent the position and physicochemical properties of the sidechains of these five residues. For a given pharmacophore, we compared each docked compound and determined whether it had a chemical moiety that overlaps with the pharmacophore with the appropriate physicochemical property. For example, a compound that possesses a benzene group that occupies the same position as an aromatic ring (pharmacophore) of a tyrosine residue on uPA is expected to mimic the properties of the tyrosine residue in the interaction between uPAR•uPA and disrupt binding. In the pharmacophore model, the ϵ -amine on Lys-23 was modeled using a positive charge feature, while the benzene rings of Tyr-24 and Phe-25 were modeled using aromatic ring features. We assigned separate aromatic ring features to the benzene and pyrrole rings on the indole of Trp-30. The aliphatic side chain of Ile-28 was represented by a hydrophobic feature centered at the C β -C γ_1 bond. We searched for compounds containing functional moieties that matched a corresponding pharmacophore feature. There were 31387, 3157, and 11 compounds in ChemDiv that overlap with one, two, and three distinct hot spots on uPA, respectively (**Table 4.2**). For DOS, fewer compounds were found to overlap with hot spots. In total, 19210, 952, and 10 compounds overlapped with one, two, and three hot spots, respectively (**Table 4.3**). The number of overlaps for SCUBIDOO was similar to DOS, with 17992, 1033, and 8 compounds that overlapped with one, two, and three hot spots, respectively (**Table 4.4**).

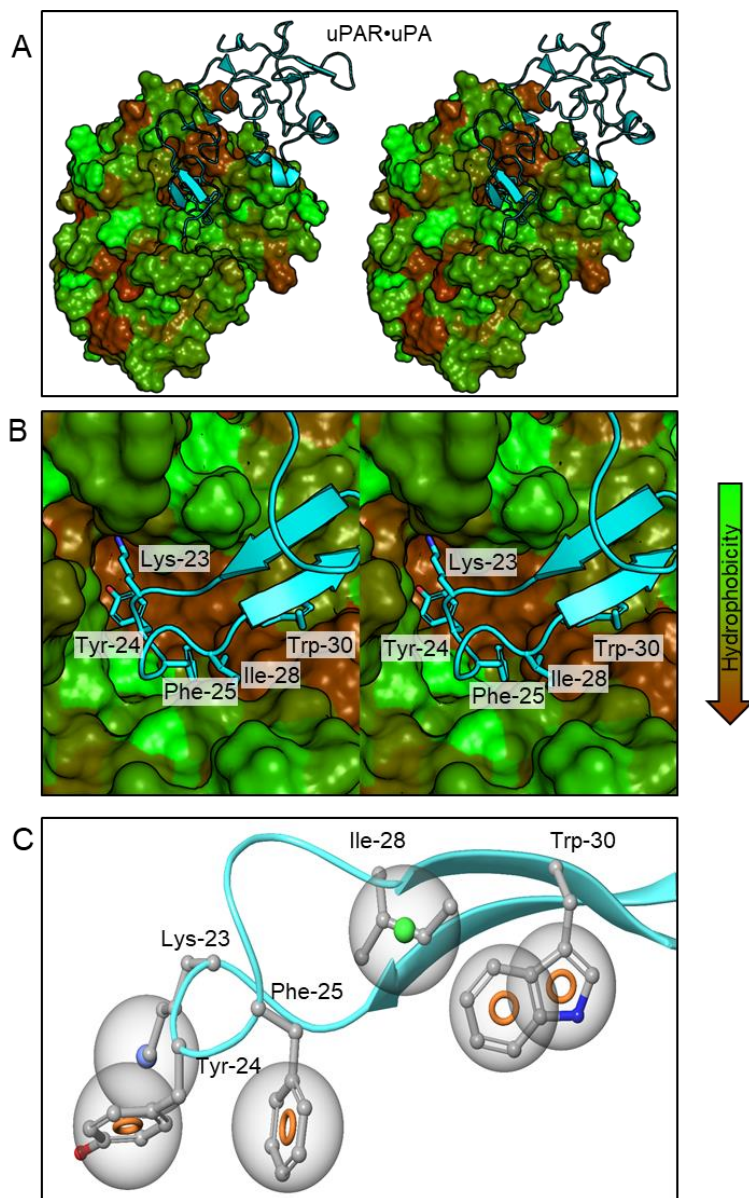


Figure 4.4. The protein complex of uPAR and uPA peptide. **(A)** uPAR is shown as a surface colored by hydrophobicity, with more hydrophobic residues in brown and more hydrophilic residues in green. The uPA peptide is shown as a cyan ribbon. **(B)** The uPAR•uPA complex and the five hot spots (sidechains in stick representation) on uPA (cyan ribbon). **(C)** Pharmacophore features for uPA are shown as small colored spheres: positive charge (Lys-23 in dark blue), hydrophobic (Ile-28 in green), and aromatic rings (Tyr-24, Phe-25, Trp-30 in tan).

Table 4.2. Matching count for compounds in ChemDiv against hot spots on uPA.

ChemDiv	Whole Collection			Top 1000 (<500 Da)		
	1	2	3	1	2	3
Lys-23	72	11	0	12	1	0
Tyr-24	13370	3118	11	64	100	0
Phe-25	17161	3040	11	427	99	0
Ile-28	784	145	11	0	0	0
Trp-30	0	0	0	0	0	0
Total	31387	3157	11	503	100	0

Table 4.3. Matching count for compounds in DOS against hot spots on uPA.

DOS	Whole Collection			Top 1000 (<500 Da)		
	1	2	3	1	2	3
Lys-23	1094	161	2	13	6	0
Tyr-24	8579	750	7	104	31	0
Phe-25	6304	541	6	277	35	0
Ile-28	1674	246	8	2	0	0
Trp-30	1559	206	7	1	0	0
Total	19210	952	10	397	36	0

Table 4.4. Matching count for compounds in SCUBIDOO against hot spots on uPA.

SCUBIDOO	Whole collection			Top 1000 (<500 Da)		
	1	2	3	1	2	3
Lys-23	773	184	8	97	64	4
Tyr-24	7618	954	8	38	85	4
Phe-25	9531	918	8	361	61	4
Ile-28	70	10	0	0	0	0
Trp-30	0	0	0	0	0	0
Total	17992	1033	8	496	105	4

There are differences in the degree of overlap with individual hot spots among the three collections. At Lys-23, DOS and SCUBIDOO compounds showed dramatically greater overlap compared with ChemDiv. A total of 1094 and 773 compounds from DOS and SCUBIDOO, respectively, overlapped with Lys-23 compared to only 72 for ChemDiv. Similarly, for compounds that overlapped with Lys-23 and one other hot spot, we found 161 DOS and 184 SCUBIDOO compounds, compared with only 11 in ChemDiv. This finding might be attributed to the larger average AlogP value of ChemDiv ($AlogP = 3.6 \pm 1.6$), which contains fewer polarizable moieties than DOS ($AlogP = 2.3 \pm 1.6$) and SCUBIDOO ($AlogP = 1.3 \pm 1.8$). Analogous trends were observed for both Ile-28 and Trp-30. At Ile-28, 784 ChemDiv and 1674 DOS compounds overlapped with the residue, while only 70 SCUBIDOO compounds showed overlap with this residue. At Trp-30, none of the compounds in ChemDiv or SCUBIDOO overlapped with either the indole or benzene rings of the sidechain, while a total of 1772 compounds from DOS overlapped with Trp-30, including 213 compounds that overlap Trp-30 and at least one other residue. The latter finding may be attributed to the fact that Trp-30 is positioned outside of the deep binding pocket on uPAR and may be more difficult to reach by the smaller compounds in ChemDiv and SCUBIDOO. When comparing the DOS compounds that overlap with Trp-30 to all DOS compounds, there was a slight shift in the distribution of aromatic rings from 4.0 ± 0.9 to 4.3 ± 0.9 , the distribution of rotatable bonds from 6.8 ± 2.1 to 7.0 ± 2.1 , and the distribution of chiral centers from 3.6 ± 0.8 to 3.7 ± 0.7 . Compared to DOS and SCUBIDOO, the number of compounds in ChemDiv that overlapped with Tyr-24 and Phe-25 was nearly double when considering only compounds that overlap with a single hotspot. Similarly, there was more than a threefold increase in the number of compounds that overlapped with Tyr-24 or Phe-25 and one other hot spot. Although DOS compounds have proven successful in producing probe molecules after high-throughput screening, many of these compounds have high molecular weight [200]. Indeed, in our study, compounds from DOS that overlap with hot spots had higher molecular weight ($MW = 536 \pm 97$ Da) than those from ChemDiv ($MW = 408 \pm 71$ Da) and SCUBIDOO ($MW = 332 \pm 45$ Da).

A question of interest is whether compounds that are predicted by docking studies to have higher binding affinities will show different overlap with hot-spot residues at the protein-protein interface. To address this question, we repeated the above analysis except that only compounds (i) that are 500 Da or less, and (ii) that are among the top 1000 highest-scoring from a docking simulation were considered. There were 603, 433 and 605 compounds from ChemDiv, DOS, and SCUBIDOO that matched at least one hot spot. ChemDiv and SCUBIDOO compounds all possessed substituents that occupy the same position as Lys-23, Tyr-24 or Phe-25 sidechains on uPA, while none overlapped with hot spots on uPA located at Ile-28 and Trp-30. The DOS

collection had only three compounds that overlapped these residues. Furthermore, except for SCUBIDOO, none of the compounds in ChemDiv or DOS simultaneously showed an overlap with three hot spots. The four compounds from SCUBIDOO overlapped with Lys-23, Tyr-24, and Phe-25. It is interesting that compounds from SCUBIDOO, which are generally smaller in size than DOS and ChemDiv compounds, are able to mimic more hot spots at the protein-protein interface.

4.2.4 TEAD•Yap. The PPI between TEAD and Yap occurs over a large interface of 1300 Å². [205] The TEAD-binding domain of Yap wraps around the globular structure of TEAD via three interfaces (**Fig. 4.5A**). The twisted-coil region of Yap at the Ω-loop is most critical for complex formation (**Fig. 4.5B**). There are six hot spots at this region, with four hydrophobic residues (Met-86, Leu-91, Phe-95, and Phe-96), one charged residue (Arg-89), and one polar residue (Ser-94). The aliphatic sidechains of Met-86 and Leu-91 were identified as hydrophobic features in the pharmacophore model, the sidechain of Arg-89 was modeled using a positive charge feature, and the benzene rings of Phe-95 and Phe-96 were modeled using aromatic ring features. The hydrogen and oxygen atom on the sidechain of Ser-94 were identified as a hydrogen-bond donor and acceptor, respectively (**Fig. 4.5C**). In the native complex, two hydrogen bonds are formed between Ser-94 on Yap and Glu-240 and Tyr-406 on TEAD [206].

There were 47061, 8862, 403, and 5 compounds in ChemDiv that overlapped with one, two, three, and four different hot spots on Yap, respectively (**Fig. 4.5D**). From the DOS collection, 50976, 9570, 425, and 3 compounds overlapped with one, two, three, and four different hot spots on Yap, respectively (**Fig. 4.5E**). From the SCUBIDOO library, only 35369, 4341, and 59 compounds overlapped with one, two, and three hot spots simultaneously, while no compounds matched four hot spots (**Fig. 4.5F**). Among compounds that overlapped with multiple hotspots, ChemDiv and DOS had similar distributions while SCUBIDOO had approximately half the number of compounds that overlapped with two hotspots and approximately 15% of the number of compounds that overlapped with three hotspots when compared to the other two collections.

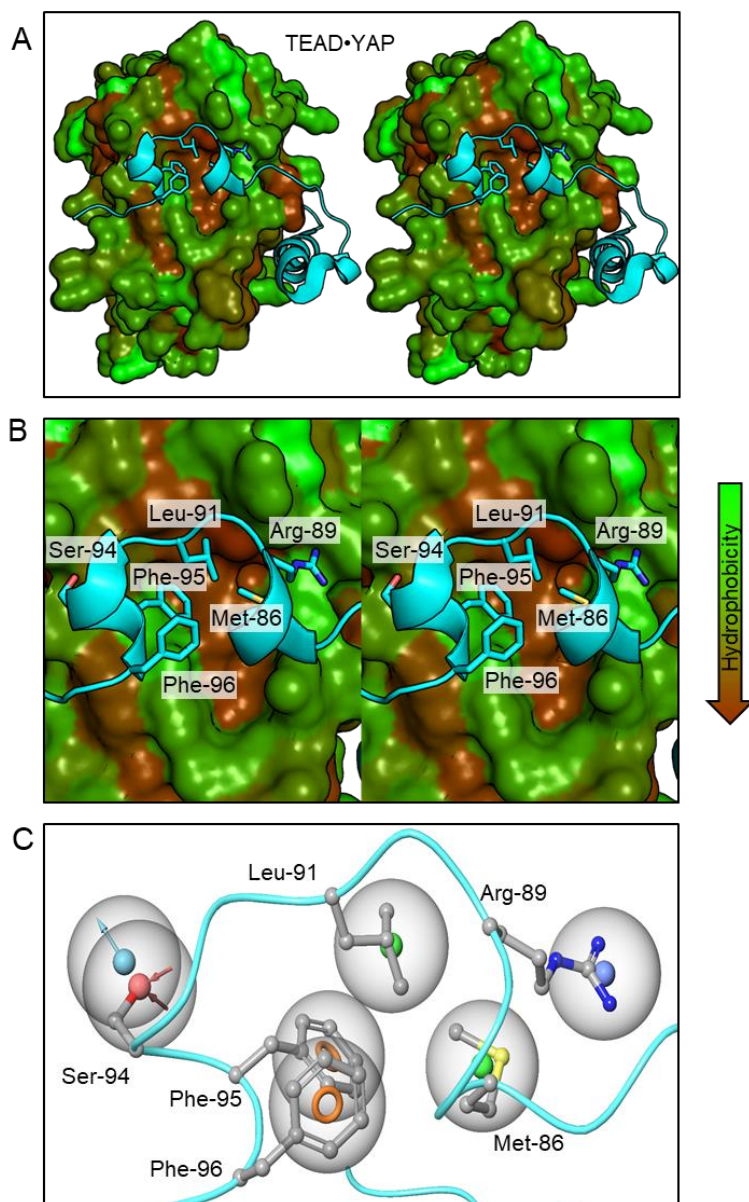


Figure 4.5. The protein complex of TEAD and YAP peptide. **(A)** TEAD is shown as a surface colored by hydrophobicity, with more hydrophobic residues in brown and more hydrophilic residues in green. The YAP peptide is shown as a cyan ribbon. **(B)** The TEAD•YAP complex and the six hot spots (sidechains in stick representation) on YAP (cyan ribbon). **(C)** Pharmacophore features for YAP are shown as small colored spheres: hydrophobic (Met-86 and Leu-91 in green), positive charge (Arg-89 in dark blue), hydrogen bond acceptor (Ser-94 in red), hydrogen bond donor (Ser-94 in light blue), and aromatic rings (Phe-95, Phe-96 in tan).

Table 4.5. Matching count for compounds in ChemDiv against hot spots on Yap.

ChemDiv	Whole Collection				Top 1000 (<500 Da)			
	1	2	3	4	1	2	3	4
Met-86	25855	7869	395	5	108	16	0	0
Arg-89	28	6	0	0	6	2	0	0
Leu-91	9890	4336	381	5	43	9	0	0
Ser-94	84	81	23	4	1	0	0	0
Phe-95	11108	5372	393	5	73	17	0	0
Phe-96	96	60	17	1	0	0	0	0
Total	47061	8862	403	5	231	22	0	0

Table 4.6. Matching count for compounds in DOS against hot spots on Yap.

DOS	Whole Collection				Top 1000 (<500 Da)			
	1	2	3	4	1	2	3	4
Met-86	29112	7746	399	3	117	83	6	0
Arg-89	2099	881	97	1	26	9	1	0
Leu-91	9501	3459	305	3	46	14	3	0
Ser-94	5481	3112	229	3	160	97	5	0
Phe-95	4580	3704	224	2	41	49	3	0
Phe-96	203	238	21	0	0	0	0	0
Total	50976	9570	425	3	390	126	6	0

Table 4.7. Matching count for compounds in SCUBIDOO against hot spots on Yap.

SCUBIDOO	Whole Collection				Top 1000 (<500 Da)			
	1	2	3	4	1	2	3	4
Met-86	23848	4180	59	0	152	38	0	0
Arg-89	89	17	3	0	6	1	0	0
Leu-91	5674	2190	55	0	89	20	0	0
Ser-94	0	0	0	0	0	0	0	0
Phe-95	5735	2283	58	0	107	19	0	0
Phe-96	23	12	2	0	0	0	0	0
Total	35369	4341	59	0	354	39	0	0

In all three sets, approximately a quarter of all compounds overlapped with the hydrophobic pharmacophore on Met-86. Similarly, about 10% in each collection overlapped with the hydrophobic Leu-91 and aromatic Phe-95 residues on Yap. At Phe-96, DOS had substantially more compounds that overlapped with the residue either alone or with one other residue compared to the other two collections. Overlap with Arg-89, Ser-94, and Phe-95 were less prevalent in ChemDiv and SCUBIDOO compared to DOS. Moreover, no compounds in SCUBIDOO matched the pharmacophore of Ser-94. This observation may be attributed to the lower molecular weight and more limited structural diversity in this library, which limits the ability of compounds to overlap with residues that are outside the immediate binding pocket.

As with uPAR•uPA, we explored hot-spot overlap of compounds that were predicted by docking to have the highest binding affinities to the protein receptor, in this case TEAD4. A collection of compounds was created by selecting 1000 compounds with the highest binding scores and molecular weights less than 500 Da. In ChemDiv, there were 231 and 22 compounds that overlapped with one or two hot spots, respectively (**Table 4.5**). In DOS, there were 390, 126, and 6 compounds that match one, two, and three hot spots, respectively (**Table 4.6**). From SCUBIDOO, we found a total of 354 and 39 compounds that matched the sidechain of one or two hot spots, respectively (**Table 4.7**). The compounds in ChemDiv and SCUBIDOO mainly mapped onto the hydrophobic residues of Met-86, Leu-91, and Phe-95, while DOS compounds generally had overlap with Ser-94 and Met-86. Interestingly, no compound overlapped with Phe-96 from any of the three collections. Moreover, in DOS, there were six compounds that overlapped with three hot spots simultaneously. All six compounds overlapped with Met-86 and have 6.8 ± 1.9 rotatable bonds and 3.8 ± 1.7 chiral centers. One compound overlapped with the nearby Arg-89, which is located at the periphery of the binding pocket. This compound is structurally complex (RB = 7, chiral centers = 3), but its molecular weight is only 345.3 Da, which suggests that compound flexibility may play a larger role than molecular weight in targeting the flat TEAD•Yap pocket.

4.2.5 Ca_vα•Ca_vβ. The structural basis of the Ca_vα•Ca_vβ interaction was revealed in a co-crystal structure between Ca_vβ₃ and the α-interacting domain (AID) of Ca_vα (Ca_vα_{AID}) (**Fig. 4.6A**). Ca_vα_{AID} is a 25-residue α-helix that binds to a well-defined groove on the GK domain of Ca_vβ₃ (**Fig. 4.6B**). Previous biophysical studies combined with site-directed mutagenesis on Ca_vα_{AID} revealed the presence of three hot-spot residues on the α-helix at Tyr-437, Trp-440, and Ile-441 [207]. The side chains of these residues are ensconced into three sub-sites. The aliphatic side chain of Ile-441 was identified as a hydrophobic feature in the pharmacophore model, while the benzene ring of Tyr-437 was modeled using an aromatic ring feature. We assigned separate aromatic ring features to the benzene and pyrrole rings on the indole of Trp-440 (**Fig. 4.6C**).

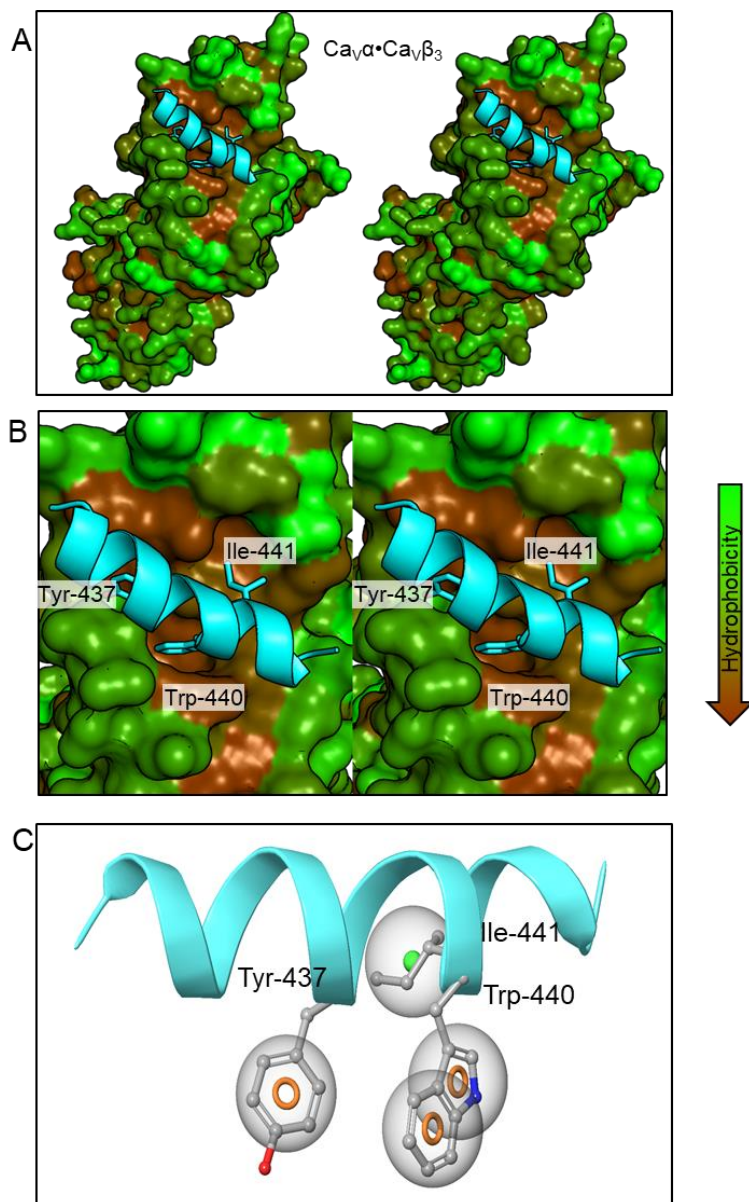


Figure 4.6. The protein complex of $\text{Ca}_v\beta$ and $\text{Ca}_v\alpha$ peptide. (A) $\text{Ca}_v\beta_3$ is shown as a surface colored by hydrophobicity, with more hydrophobic residues in brown and more hydrophilic residues in green. The $\text{Ca}_v\alpha$ peptide is shown as a cyan ribbon. (B) The $\text{Ca}_v\alpha \cdot \text{Ca}_v\beta_3$ complex and the three hot spots (sidechains in stick representation) on $\text{Ca}_v\alpha$ (cyan ribbon). (C) Pharmacophore features for $\text{Ca}_v\alpha$ are shown as small colored spheres: hydrophobic (Ile-441 in green) and aromatic rings (Tyr-437, Trp-440 in tan).

Table 4.8. Matching count for compounds in ChemDiv against hot spots on Ca_vα.

ChemDiv	Whole Collection			Top 1000 (<500 Da)		
	1	2	3	1	2	3
Tyr-437	26800	33684	2099	97	509	50
Trp-440	28235	35170	2099	226	529	50
Ile-441	4136	9026	2099	10	68	50
Total	59171	38940	2099	333	553	50

Table 4.9. Matching count for compounds in DOS against hot spots on Cav α .

DOS	Whole Collection			Top 1000 (<500 Da)		
	1	2	3	1	2	3
Tyr-437	20509	7980	476	270	104	2
Trp-440	25864	9767	476	327	90	2
Ile-441	11846	8953	476	62	108	2
Total	58219	13350	476	659	151	2

Table 4.10. Matching count for compounds in SCUBIDOO against hot spots on $\text{Ca}_{\nu\alpha}$.

SCUBIDOO	Whole Collection			Top 1000 (<500 Da)		
	1	2	3	1	2	3
Tyr-437	19108	12144	40	99	478	2
Trp-440	22112	15168	40	301	544	2
Ile-441	3877	4202	40	2	72	2
Total	45097	15757	40	402	547	2

There were 59171, 38940, and 2099 compounds in ChemDiv that overlapped with one, two, and three hot spots, respectively (**Table 4.8**). From DOS and SCUBIDOO, substantially fewer compounds were found to overlap with multiple hot spots. There were 58219, 13350, and 476 compounds from DOS that overlapped with one, two, and three separate hot spots, respectively (**Table 4.9**), and correspondingly 45097, 15757, and 40 from SCUBIDOO (**Table 4.10**). While the number of compounds that overlap with a single hot spot are similar, the number of compounds that overlap with two or three hot spots is substantially different for ChemDiv compared to DOS and SCUBIDOO. The largest difference was seen in compounds that overlap with either Tyr-437 or Trp-440 as well as one other hot spot. In ChemDiv, approximately 35000 compounds overlapped with one of the two aromatic hot spots as well as another residue. If a compound in ChemDiv was overlapping with multiple residues, it was often at both the Tyr-437 and Trp-440 sites, while fewer compounds overlapped with one of these two aromatic residues and the hydrophobic Ile-441. This trend was also observed in SCUBIDOO, where there was much higher co-occurrence of a combination of Tyr-437 and Trp-440 than between one of the two aromatic residues and Ile-441. Among DOS compounds, in contrast, the co-occurrence rate between the three residues was more uniform, with 7980, 9767, and 8953 compounds that overlapped with Tyr-437, Trp-440, and Ile-441 and one other residue, respectively.

We again considered compounds with highest predicted binding affinities and low molecular weight. The top highest-scoring 1000 compounds with molecular weights less than 500 Da were selected based on docking score. There were 936, 812, and 951 compounds for ChemDiv, DOS, and SCUBIDOO matching at least one hot spot. It is interesting that despite the higher complexity of DOS compounds (chiral centers = 3.5 ± 1.0) fewer matched hot spots than ChemDiv and SCUBIDOO compounds, which are less rich in chiral centers (ChemDiv chiral centers = 0.4 ± 0.7 ; SCUBIDOO chiral centers = 1.3 ± 1.0). Interestingly, there were twice as many compounds in ChemDiv and SCUBIDOO that overlapped with at least two hot spots than DOS. Remarkably, there were 50 compounds in ChemDiv that overlapped with three hot spots compared to only two each for DOS and SCUBIDOO.

4.2.6 Virtual Screening of Commercial Library Against Two Protein-Protein Interactions. The discovery that conformationally-restricted fragment-like libraries such as SCUBIDOO can be effective in mimicking protein-ligand hot spots for interfaces with well-defined pockets prompted us to test this observation with virtual screening and experimental validation. The ChemDiv commercial library was filtered for conformationally-restricted compounds and tested against the uPAR•uPA interaction. The ChemDiv library was filtered for compounds between 350 and 500 Da having six or fewer rotatable bonds as well as high predicted solubility

(AlogP \leq 4). The resulting collection of 50,893 compounds was docked to uPAR at the uPAR•uPA interface and to TEAD at the TEAD•Yap interface. Compounds were ranked based on their overlap with hot-spots located at the protein-protein interface. The top 85 compounds were selected and tested for inhibition of uPAR•uPA and TEAD•Yap using a previously developed fluorescence polarization assay. The compounds were first tested in duplicate at a single concentration of 50 μ M (**Fig. 4.7**). For the uPAR•uPA interface, five compounds, namely **1** (IPR-3247), **2** (IPR-3260), **3** (IPR-3271), **4** (IPR-3288) and **5** (IPR-3305) inhibited more than 25%. These compounds were also tested in a concentration-dependent manner (**Fig. 4.8**). All five compounds inhibited the FP assay, with K_i s in the double-digit micromolar range. All five were predicted to overlap with two aromatic hot spots on uPA, Tyr-24 and Phe-25 (**Fig. 4.9**). Compound **1**, however, appears to have poor solubility at concentrations that are higher than 25 μ M. Compounds **3** and **5** also exhibited poor solubility at higher concentrations. The two compounds with reasonably good solubility were **1** and **4** with K_i s of 25 and 62 μ M, respectively.

Conformationally-restricted compounds from ChemDiv were docked against TEAD at the TEAD4•Yap interface. The top 81 compounds from virtual screening against the Ω -loop pocket on TEAD4 were experimentally validated using a fluorescence polarization assay. Compounds were tested in duplicate at an initial concentration of 50 μ M. No compounds were found to inhibit more than 15%. TED-97, a previously discovered small-molecule inhibitor of TEAD4•Yap1 was used as positive control. A concentration-dependent study was performed for compounds that inhibited by 15%, but none exhibited concentration-dependent inhibition. The lack of hits confirms the highly challenging nature of this interaction and the fact that to successfully inhibit this target, larger more complex compounds such as DOS may be required.

The ranking of the filtered and unfiltered compound sets from uPAR was compared to determine whether pre-filtering by rotatable bonds, molecular weight, and AlogP led to further enrichment of the original library for hit compounds. We identified the ranking of the five active compounds using the distribution of dockingscores of the compounds in the unfiltered ChemDiv library (**Fig. 4.10**). For the unfiltered library, we found that four of the five hit compounds have docking scores that puts them between the 25th and 75th percentiles. In other words, had we strictly ranked the unfiltered library, these hit compounds would not have been selected as they would not have been among the top candidates chosen for experimental validation. Only compound **3** (IPR-3271) would have been among the top 100 compounds in the unfiltered library (i.e., without physicochemical filters).

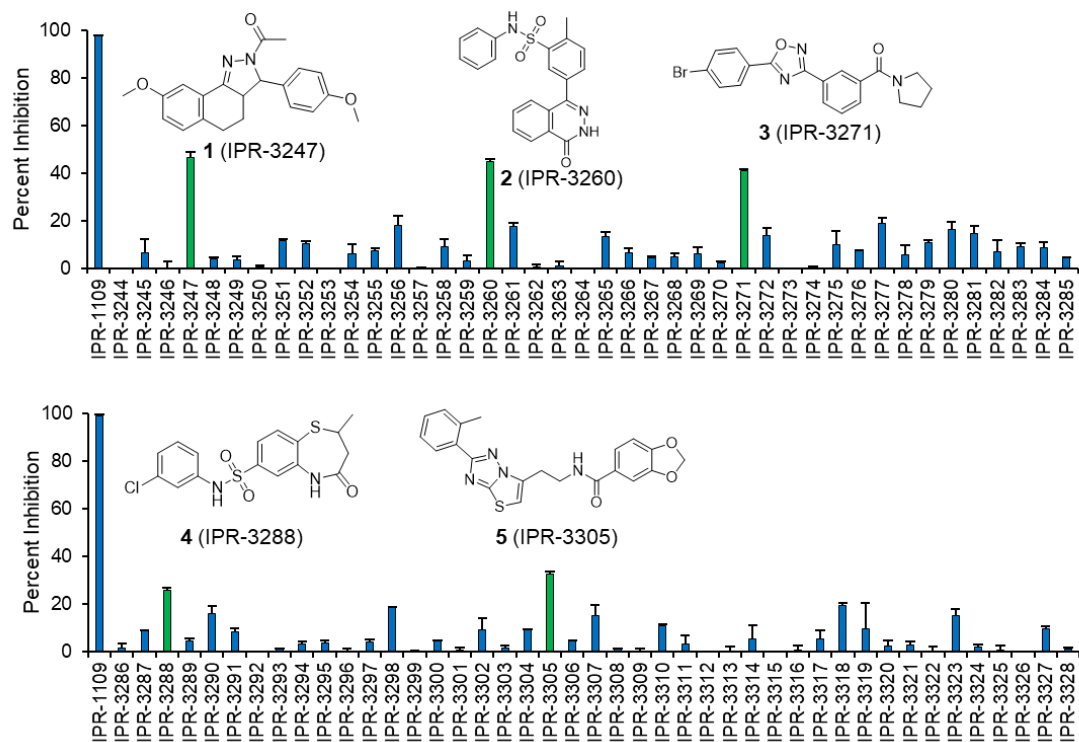


Figure 4.7. Virtual screening of commercial library against uPAR. A set of 85 compounds from virtual screening were tested at a single concentration of 50 μ M in a fluorescence polarization (FP) assay for inhibition of uPAR•AE147-FAM interaction. Structures are provided for compounds **1-5** (green bars) that were advanced to concentration-response experiments.

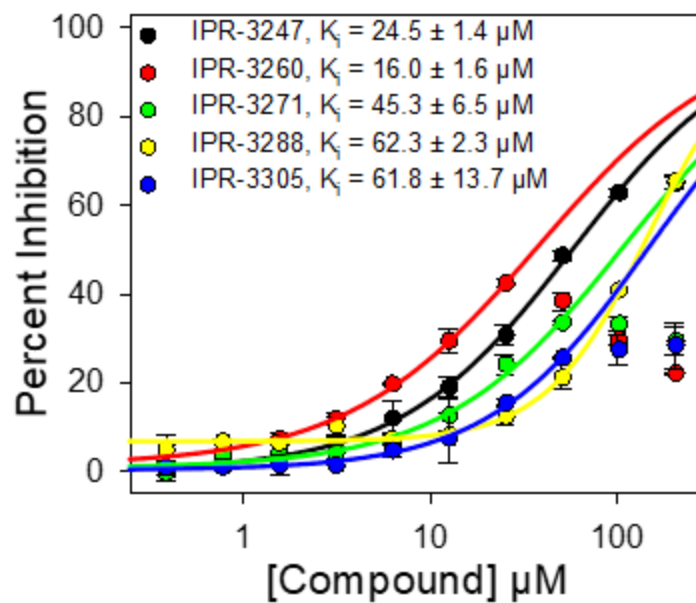


Figure 4.8. Concentration-dependent of compounds identified from initial screening against uPAR. Concentration-dependent FP assay measuring the inhibition of uPAR•AE147-FAM peptide interaction by compounds. At high concentrations, some compounds were insoluble and high-concentration data points were omitted from curve-fitting.

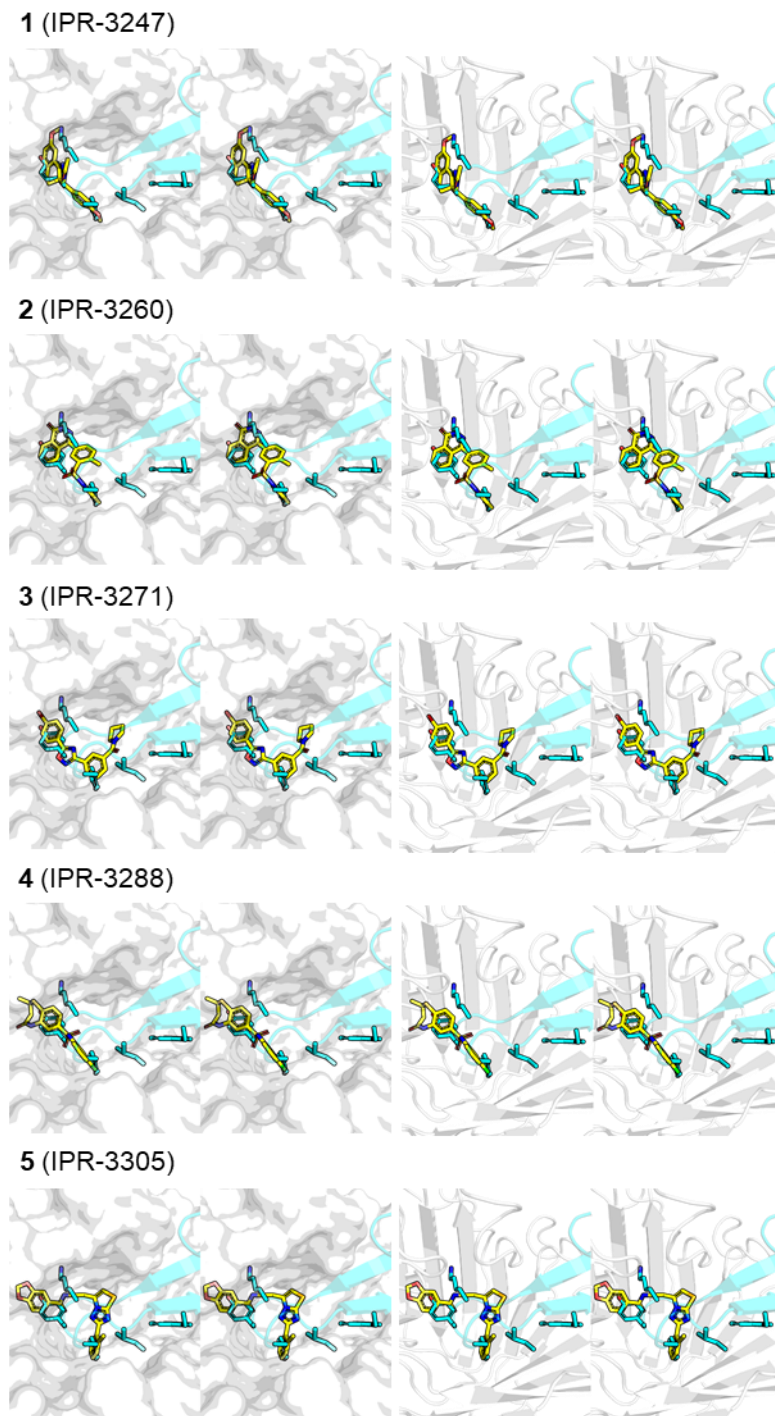


Figure 4.9. Binding modes of hit compounds. Virtual screening binding modes of **1-5** (yellow sticks) in the uPAR•uPA binding pocket. uPAR is shown as a white surface and ribbons on the left and right, respectively. uPA is shown as a partially transparent cyan ribbon. The side chains of four interface residues on uPA are shown as a stick representation and colored cyan.

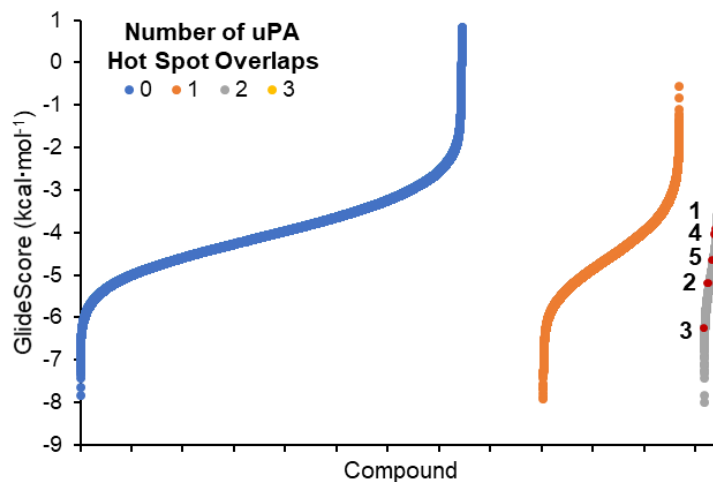


Figure 4.10. Distribution of docking scores (GlideScores) for all compounds in ChemDiv compared to the five active compounds identified. All compounds in ChemDiv are grouped by the number of uPA hot spots overlapped and rank-ordered by GlideScore. Five active compounds (red dots) are depicted: **1** (IPR-3247), **2** (IPR-3260), **3** (IPR-3271), **4** (IPR-3288), and **5** (IPR-3305).

4.3 DISCUSSION

Three chemical libraries were compared for their ability to mimic the positions of amino-acid sidechains of critical residues at protein-protein interfaces: (i) a commercial library from ChemDiv; (ii) a collection of diversity-oriented synthesis (DOS) libraries from the Broad Institute; and (iii) SCUBIDOO, a non-commercial library built by combining building blocks of common reactions in medicinal chemistry. The physicochemical properties of each of these three collections was explored. DOS compounds were larger, had more chiral centers, and were more flexible as evidenced by a larger number of rotatable bonds compared with ChemDiv and SCUBIDOO. Similarly, the molecular weight distributions of the compound collections correlate well with the number of rotatable bonds and chiral centers in DOS and ChemDiv, but not SCUBIDOO. Principal component analysis revealed a primary separation of all three collections in the first component, while the second principal component did not distinguish DOS from SCUBIDOO libraries. Finally, principal moment of inertia analysis suggests that compounds in the SCUBIDOO library are enriched for linear compounds, while ChemDiv and DOS contain compounds with more disc-like and some globular characteristics.

We explored the ability of these three compound collections to mimic the physicochemical properties of amino acid side chains of hot spots at the interface of three PPIs. Three tight PPIs were selected: (i) a β -turn motif at the interface between the urokinase receptor and its serine proteinase ligand urokinase (uPAR•uPA), (ii) a twisted-coil motif at the Ω -loop between the transcription factor TEAD and its co-activator Yap (TEAD•Yap), and (iii) an α -helix motif between the α - and β -subunits of the calcium channel ($\text{Ca}_v\alpha\text{•Ca}_v\beta$). In each system, hot-spot residues were identified on the protein ligand at each PPI interface. Each library showed different overlap with individual hot spots in each PPI. For example, for uPAR•uPA, ChemDiv contains more compounds that overlap with the two aromatic residues Tyr-24 and Phe-25, while DOS and SCUBIDOO possess many more compounds that overlap with the positively charged Lys-23 residue. Similarly, DOS was twice as likely to overlap with Ile-28 than ChemDiv and 20 times as likely than SCUBIDOO. DOS was also the only collection to have compounds that overlapped with the more distant Trp-30 hot spot. Similar trends were observed in TEAD•Yap, where DOS compounds were better able to mimic Arg-89, Ser-94, and Phe-96 compared to ChemDiv and SCUBIDOO, but ChemDiv compounds were better able to mimic Phe-95 compared to the other two collections. In $\text{Ca}_v\alpha\text{•Ca}_v\beta$, we found similar distributions in the number of compounds that overlapped with a single hot spot, but more than two and four times the number of compounds that overlapped with two and three distinct hot spots in ChemDiv compared to DOS. The gap is further exacerbated

when comparing ChemDiv and SCUBIDOO, where there were two and 50 times as many compounds that overlapped with two and three hot spots, respectively.

We refined our analysis, except that the top 1000 best scoring compounds by docking simulation were selected, subject to a molecular weight filter. This new set of compounds was enriched for small molecules predicted to bind to the target with high affinity. Interestingly, there were many more small molecules from ChemDiv and SCUBIDOO that overlapped with at least two hot spots for uPAR•uPA and $Ca_v\alpha\bullet Ca_v\beta$. In fact, for the $Ca_v\alpha\bullet Ca_v\beta$ interaction, there were more than 50 compounds in ChemDiv that overlapped with three hot spots, compared with only two for both DOS and SCUBIDOO. For TEAD4•Yap1, the DOS collection showed substantially greater overlap. These results can be explained by the fact that when compounds are ranked by their binding affinity, the resulting set is likely enriched for small molecules that make the best shape complementarity with existing pockets and therefore result in the largest solvent-accessible surface area change. In the case of uPAR and $Ca_v\beta$, both receptors possess well-defined binding pockets. Considering that ChemDiv and SCUBIDOO possess a large number of small fragment-like compounds, it is more likely that these compounds will fit best into pockets and sub-pockets of these receptors. DOS compounds, on the other hand, are larger and more flexible, and may not score as well as SCUBIDOO and ChemDiv library small molecules, which may explain why there are fewer of these compounds that overlap with two or more hot spots in the top 1000 set. Since TEAD4 has a shallow pocket, DOS compounds will probably score higher, since for a flat surface, a larger compound is much more likely to lead to greater surface burial and therefore a better score. These results were tested by carrying out a virtual screen of more than 50,000 compounds from the ChemDiv library against uPAR and TEAD. We selected conformationally-restricted compounds between 350 and 500 Da in MW. Interestingly, several hits emerged in uPAR, and the most promising among them are fragment-like hits that have 2-3 rotatable bonds. No hits were identified for TEAD4.

Our work suggests that small fragment-like compounds that are conformationally-restricted can be good candidates for PPI with well-defined pockets, while larger more complex compounds may be a better option for PPI with large but flat and featureless surfaces. The affinity of large hit compounds is generally driven by entropy, so these compounds may not show optimal fit to smaller, well-defined binding pockets. Hence, improving the enthalpy of binding becomes a significant challenge that is only accomplished by substantial modifications to the compound structure. Small fragment-like compounds are generally driven by enthalpy. These compounds tend to show optimal fit into pockets. It is easier to modify these compounds by extending them into neighboring pockets while at the same time improving entropy of binding.

4.4 MATERIALS AND METHODS

4.4.1 Ligand Preparation. Three libraries were chosen to compare how compounds mimicked the hot-spot residues on the different PPI interfaces. The first is the ChemDiv library, which consists of over 1.7 million compounds that were retrieved from the ZINC15 website [208]. The second collection is a set of diversity-oriented synthesis (DOS) libraries prepared at the Broad Institute. This collection consists of 100,903 compounds from 32 libraries that were synthesized using DOS principles applied to 11 different reaction pathways [194, 200]. The third collection is from SCUBIDOO, which was developed from 58 organic reactions known in the pharmaceutical field totaling over 21 million compounds [199]. The SCUBIDOO website provides three different representative samples created using a stratified sampling algorithm. We selected the M library from SCUBIDOO consisting of 99,977 compounds. Compounds predicted to be pan-assay interference compounds (PAINS) were identified and filtered using the PAINS1, PAINS2, and PAINS3 filters in Canvas [209, 210], resulting in 1,428,800 compounds for ChemDiv, 99,663 compounds for DOS, and 93,074 compounds for SCUBIDOO. ChemDiv compounds retrieved from ZINC were previously prepared through their internal workflow [211]. Compounds from DOS and SCUBIDOO were prepared using LigPrep [173]. Epik was used for protonation-state assignment and tautomer generation [125]. The OPLS_2005 force field was used for minimization and the ionization states were generated at pH 7 [212]. Compounds were desalted to exclude additional molecules such as counter ions in salt and water molecules and tautomers were generated. For DOS and SCUBIDOO, stereoisomers were generated by retaining specified chiralities and varying those where the stereochemistry of the chiral center were undefined. Up to 32 different stereoisomers per ligand were generated in this manner, resulting in 127,483 compounds for DOS and 125,917 compounds for SCUBIDOO. The size of each compound library was normalized by random sampling to 125,917 compounds.

4.4.2 Principal Component Analysis. Principal component analysis (PCA) aims to simplify high-dimensional data by projecting the data onto a new set of dimensions that most effectively captures the variance in the data. We used it to visualize similarities and differences between the physicochemical properties of different collections of compounds. Eight physicochemical features were used for PCA: molecular weight (MW), aLogP, number of hydrogen-bond acceptors (HBA), number of hydrogen-bond donors (HBD), number of rotatable bonds (RB), polar surface area (PSA), number of chiral centers, and number of rings. The mean of each feature is shifted to zero and each feature is scaled to have unit variance prior to the analysis. PCA is calculated using singular value decomposition in R (version 3.2.3).

4.4.3 Principal Moment of Inertia. Principal moment of inertia (PMI) descriptors provide an intuitive notion of the three-dimensional shape diversity of the various compound data sets. Low-energy conformations are identified for each molecule in the data set, and three PMI values (I_1 , I_2 and I_3 ; where $I_3 \geq I_2 \geq I_1$) are calculated for each conformer. Normalized ratios of PMI (I_1 / I_3 and I_2 / I_3) are then calculated and plotted on a triangular graph, with the vertices (0,1), (0.5,0.5), and (1,1) representing a perfect linear (diacetylene), disc-like (benzene), and spherical (adamantane) compound, respectively. The moments of inertia in the three spatial dimensions were calculated by the *calculate_pmi.py* script in Schrödinger.

4.4.4 Protein Preparation. The structures of the uPAR•uPA (PDB ID: 3BT1), TEAD•Yap (PDB ID: 3KYS), and Cav α •Cav β (PDB ID: 1VYT) interactions were retrieved and prepared using the Protein Preparation Wizard using the Schrödinger Suite [173, 213]. Bond orders were assigned, hydrogen atoms were added, and disulfide bonds were created. Water residues and additional ions and heteroatom groups were discarded. Missing sidechains and loops were introduced using the Prime module [123]. The resulting protein structures were protonated at pH 7.0 using PROPKA [124].

4.4.5 Virtual Screening. The compound library was docked to the prepared protein structures using AutoDock Vina (Version 1.1.2) [172]. The binding pocket was centered at each of the interfaces with a box with dimensions of 21 Å × 21 Å × 21 Å. All other parameters were set to default values. The docked conformations were converted back to MOL2 format using in-house Python scripts for additional analysis. Glide was used to score the Vina-docked binding modes in place using the GlideHTVS scoring function [214, 215].

4.4.6 Ligand Pharmacophore. A previously described [178] pharmacophore-based approach was adapted and used to identify how docking compounds overlapped with and mimicked known hot spots of the protein ligand in each of the protein-protein complexes. A set of pharmacophore hypotheses was constructed corresponding to the physicochemical properties of the protein ligand sidechain using the Phase package in Schrödinger [115, 116]. Phase has six built-in types of pharmacophore features: (i) hydrogen-bond acceptor (A); (ii) hydrogen-bond donor (D); (iii) hydrophobe (H); (iv) negative ionizable (N); (v) positive ionizable (P); and (vi) aromatic ring (R). The docked ligand conformation was used for pharmacophore calculations with an intersite distance matching tolerance of 2.0 Å. Hydrogen-bond acceptor sites were positioned on atoms that carry one or more donatable lone pairs, while hydrogen-bond donor sites were centered on each electrophilic site. Negative and positive ionizable sites were modeled as single points located on a formally charged atom, or at the centroid of a group of atoms over which the ionic charge is shared. Rings, isopropyl groups, t-butyl groups, various halogenated moieties, and aliphatic chains are

treated as hydrophobic sites [116]. Aromatic rings were distinguished from other hydrophobic groups and designated as a separate type of pharmacophore feature (i.e., “R” rather than “H”). In these cases, a single site was placed at the centroid of each aromatic ring. In aromatic ring pharmacophores, a normal vector was projected orthogonal to the plane of each ring. Similarly, in positive and negative ionizable pharmacophores (i.e., “P” and “N”, respectively), a vector parallel to the plane of the respective atom was formed.

Compound conformers from virtual screening were used to identify matches without refinement using Phase’s default fitness function. In this scoring function, three factors are used to describe the degree to which a compound matches a pharmacophore feature: (i) the site score, which describes how well the compound superimposes the pharmacophore feature, (ii) the vector score, which describes the cosine angle between the normal vector of an aromatic ring on the compound with an aromatic feature, and (iii) the volume score, which describes the proportion of the total volume of the pharmacophore feature overlapped by the compound. The vector score ranges from -1.0 (antiparallel) to 1.0 (parallel) in positive and negative ionizable pharmacophores, and 0.0 (perpendicular) to 1.0 (parallel) in aromatic pharmacophores. Vector scores less than 0.8, corresponding to angles more than $\pm 37^\circ$, were rejected. Compounds with either root-mean-square deviation (RMSD) overlap greater than 1.2 Å or that did not overlap with the pharmacophore feature were discarded. All other parameters were set at default values. The remaining compounds that matched a given pharmacophore were retained without sorting compounds by Phase’s internal scoring function.

4.4.7 Fluorescence Polarization (FP) Assay. Polarized fluorescence intensities were measured using EnVision Multilabel plate readers (PerkinElmer, Waltham, MA) with excitation and emission wavelengths of 485 and 535 nm, respectively [165]. Samples were prepared in Thermo Scientific Nunc 384-well black microplate in duplicates. First, the compounds were diluted in DMSO and further diluted in 1× PBS buffer with 0.01% Triton X-100 for a final concentration of 200 – 0.2 μM. Triton X-100 was added to the buffer to avoid compound aggregation. 5 μL of the compound solution and 40 μL of PBS with 0.01% Triton X-100 containing uPAR was added to the wells and incubated for at least 15 min to allow the compound to bind to the protein. Finally, 5 μL of fluorescent AE147-FAM peptide solution was added for a total volume of 50 μL in each well resulting in final uPAR and peptide concentrations of 250 nM and 100 nM respectively. For the TEAD4•Yap1 FP assay, final concentrations of 64 nM GST-TEAD4 (217-434) and 16 nM FAM-labeled Yap1 peptide (residues 60-99) were used. The final DMSO concentration was 2% v/v, which had no effect on the binding of the peptide. Controls included wells containing only the peptide, and wells containing both protein and peptide, each in duplicate to ensure the

reproducibility of the assay. When compounds were insoluble and visible precipitation was observed, the data points at high concentrations were not included in the calculation of IC_{50} values. Inhibition constants were calculated from the IC_{50} values using the K_i calculator available at http://sw16.im.med.umich.edu/software/calc_ki/.

Chapter 5

SMALL-MOLECULE BINDING SITES TO EXPLORE NEW TARGETS IN THE CANCER PROTEOME

5.1 INTRODUCTION

Large-scale sequencing studies of human tumors such as The Cancer Genome Atlas project (TCGA) provide an opportunity to uncover the genetic basis of the processes that drive cancer. Analysis of this genomic data has revealed that the complex phenotypes that define cancer are driven by tens of somatic mutations that occur on proteins across the cellular network [216]. Recent whole genome sequencing studies have profiled the molecular signatures of individual tumors to identify underlying driver mutations of each disease [4-9, 217]. Tumors were found to harbor tens of mutations. Whole-genome gene expression profiling studies have been instrumental not only in classifying tumors and uncovering genetic alterations in cancer cells (mutations, copy number, and rearrangements), but as a rich source of potential targets in cancer [2, 3]. A growing list of three-dimensional protein structures make it now possible to rationally develop small-molecule probes to explore these targets. Small-molecule probes can also provide leads for drug-discovery validation.

TCGA is an ongoing effort that aims to catalog clinical and molecular profiles of tumor samples from over 30 cancer types to discover cancer-causing alterations in large cohorts through integrated multi-platform analyses. The project aims to integrate the clinical and molecular profiles of at least 500 tumors for each disease and to determine its underlying molecular mechanism. Multiple platforms capture the clinical, pathological, genomic, epigenomic, transcriptomic, and proteomic profiles of cancers in TCGA project. Among these platforms, RNA-seq is a widely-used technology for the characterization of mRNA expression. RNA-seq uses high-throughput short reads that offer several distinct advantages over its array-based predecessors. RNA-seq is not limited by a set of predetermined probes seen in microarrays, and is superior in its ability to identify low abundance transcripts, biological isoforms, and genetic variants [218]. RNA-seq was performed for both tumor and normal tissue for each disease at TCGA. Comparison of tumor and normal mRNA levels can be used to identify overexpressed genes and their corresponding protein product that may contribute to tumor formation, progression, and metastasis. Patient information that accompanies the genomic data affords further analyses to assess the correlation of mRNA levels with patient outcome. Survival curves constructed by plotting patient outcome with time can be used to generate metrics such as hazard ratios and other coefficients to determine the correlation between overexpression of individual genes and clinical outcome. This analysis has been widely

used in clinical trials, where Kaplan-Meier survival curves are used to determine the time-to-event differences between placebo and drug groups [219].

Extensive data from TCGA combined with the exponentially growing structural data at the Protein Data Bank (PDB) offers a unique opportunity to identify protein structures of overexpressed or clinically-relevant genes in cancer. These structures can be used to scan for binding sites to develop chemical probes and lead compounds for drug discovery. In addition to detecting binding sites, algorithms have been developed to score these binding sites based on whether they can accommodate a small molecule. Both SiteMap and fpocket provide descriptors to assess binding sites that are suitable for small-molecule ligands based on the amino acid composition of the binding site and its collective physicochemical properties. SiteMap uses the hydrophobicity and accessibility of a detected binding site to assess how likely a small-molecule inhibitor will bind. It provides two scores, SiteScore and DrugScore. The latter score goes beyond just assessing a binding site for ligand binding. It measures whether a binding site is druggable, or whether it possesses similar proteins to other binding sites that have led to FDA-approved drugs. fpocket provides a measure called the Druggability Score, which is a general logistical model based on the local hydrophobic density of the binding site, as well as a hydrophobicity and normalized polarity score. The discovery of binding sites within structures that are encoded by overexpressed genes with clinical relevance is highly significant as these binding sites can be used to develop novel cancer therapeutics that are likely to exhibit greater efficacy in humans.

In addition to druggability, the binding sites must be functionally important to serve as targets for small molecules. For example, binding sites located at enzyme active sites or at the interface between a protein-protein complex are expected to disrupt protein function. Protein kinases are one example of an enzyme class with druggable binding sites that occur at the enzyme active site [220]. The ATP binding site of kinases is highly druggable with a SiteMap SiteScore and DrugScore above 1.1 [221]. There are fewer small-molecule inhibitors of protein-protein interactions, which is partly due to the lack of druggable binding sites at protein-protein interfaces. The only examples of PPI inhibitors that have shown *in vivo* efficacy, such as MDM2/p53 or Bcl-xL, possess druggable binding sites (DrugScore of 0.92 and 0.82, respectively) [222]. Therefore, the identification of binding sites that are considered druggable at protein-protein interaction interfaces can provide new avenues to develop chemical probes and cancer therapeutics. Finally, it is worth mentioning that binding sites located outside an enzyme active site or protein-protein interface can also be functionally relevant. These binding sites may modulate protein function in an allosteric manner through long-range interactions that involve dynamic changes of the target

protein [167, 223-226]. Allosteric inhibitors have been successfully used to inhibit kinase activity and in some cases, such as *AKT*, have shown more promise than competitive inhibitors.

Here, we collect gene expression profiles for 10 cancer types from TCGA and compare the expression profiles between cancer and normal samples to identify genes that are overexpressed in each cancer type. We search the Protein Data Bank for crystal structures of the protein products of these genes. We scan the surface of these proteins and identify binding sites. The functional relevance of these binding sites is explored by classifying them into known enzyme active sites, protein-protein interaction sites, or other sites that may lie outside of functional sites. To further explore the biological outcome of small molecules that bind to these binding sites, proteins harboring binding sites are further characterized in the context of a global PPI network and cancer signaling pathways to gain insight into the biological effect of binding at these binding sites. Patient data is used to investigate the correlation of overexpressed genes with clinical outcome. Our analysis uncovered new unexplored and potentially druggable and clinically-relevant protein targets. The study also provides new avenues for the rational design of small-molecule probes for well-established oncogenes. This is the first study that maps binding pockets on three-dimensional structures of the PDB within the context of cancer genomic data.

5.2 RESULTS

5.2.1 Three-Dimensional Structures of Proteins Encoded by Differentially Expressed Genes. We collected mRNA gene expression profiles of 10 cancer types from TCGA: breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), glioblastoma multiforme (GBM), head-and-neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), thyroid adenocarcinoma (THCA), triple-negative breast cancer (TNBC), and uterine corpus endometrioid carcinoma (UCEC). For each cancer type, we collected the gene expression profiles of both normal and tumor samples from RNA sequencing platforms using TCGA's Level 3 data. A search from among the 20192 reference proteins using UniProt [227] identifiers led to 7044 proteins that are encoded by TCGA overexpressed genes (**Table 5.1**).

For each cancer type, we identified the number of overexpressed genes with protein products having at least one high-resolution crystal structure by mining the Protein Data Bank. A total of 5069 unique protein chains on 2758 crystal structures from the PDB mapped to at least one of the 7044 overexpressed genes. In cases where more than one crystal structure was identified for a protein, the computer program CD-HIT was used to cluster the protein sequences of the crystal

structures to find a set of non-redundant representative structures for the given protein. This resulted in 1624 unique crystal structures of proteins encoding overexpressed genes.

The total number of proteins that encoded overexpressed genes ranged from 839 for TNBC to 2096 for LUSC (**Table 5.2**). Overall, the percentage of differentially expressed genes with at least one crystal structure spanning at least a portion of the gene sequence ranges from 20% in LUSC to 34% in GBM.

Additionally, we introduce more stringent cutoffs to distinguish between proteins that can act as probes versus those that feature druggable binding sites by increasing cutoffs of both the \log_2 fold change and the druggability property of a binding site. Using these increased cutoffs, we identify 5218 overexpressed proteins in TCGA, with only 1218 having a high-quality crystal structure at the PDB (**Table 5.1**).

5.2.2 Identification of Binding Sites on Protein Structures at the PDB. Using the three-dimensional structure of overexpressed genes for each disease, we scanned their surfaces for binding sites using the SiteMap computer program. SiteMap identifies binding sites by overlaying a three-dimensional grid around the entire protein to determine the van der Waals energies at each point of the grid (site point). By linking together site points on the protein surface that are protected from the solvent, SiteMap identifies potential binding sites on a protein surface. Each binding site identified by SiteMap is evaluated based on its ability to bind a ligand (SiteScore) and its druggability (DrugScore). Both SiteScore and DrugScore use the weighted sums of the same parameters, namely the (i) number of site points in the binding site; (ii) enclosure score that is a measure of how open the binding site is to solvents; and (iii) hydrophilic character of the binding site (hydrophilic score). Unlike DrugScore, SiteScore limits the impact of hydrophilicity in charged and highly polar sites. A binding site with SiteScore and DrugScore of 0.8 is considered to be able to fit a small molecule ligand. SiteScore and DrugScore values closer to 0.8 are considered ‘difficult’ to drug, while binding sites with SiteScore and DrugScore closer to 1.1 are classified as highly ‘druggable’ [221]. In this work, we consider a binding site with SiteScore and DrugScore values of 0.8 or greater as able to be probed and a binding site with DrugScore greater than 1.0 as druggable.

Table 5.1. Structural coverage of TCGA and the human proteome.

	TCGA Druggable Binding Sites (log₂FC ≥ 2.0, DS ≥ 1.0)	TCGA Binding Sites (log₂FC ≥ 1.5, DS ≥ 0.8)	All Proteins
Total Number of Proteins	5,218	7,044	20,192
Proteins with Structure	1,218	1,624	4,124
Proteins with Druggable Binding Sites	405	1,044	2,607
Number of Druggable Binding Sites	502	2,214	5,498
ENZ	126	434	
PPI	55	231	
OTH	331	1,576	

Table 5.2. Distribution of protein structures and druggable binding sites among cancer types ($\log_2FC \geq 2.0$, $DS \geq 1.0$).

Cancer Type	Cancer Name	Total Number of Proteins	Proteins With Structure	Proteins with Druggable Binding Sites	Number of Druggable Binding Sites	Binding Site Type		
						ENZ	PPI	OTH
BRCA	Breast invasive carcinoma	1314	280	79	93	29	14	54
COAD	Colon adenocarcinoma	971	187	47	64	15	8	45
GBM	Glioblastoma multiforme	1168	429	161	145	34	13	99
HNSC	Head and neck squamous cell carcinoma	697	128	28	34	10	4	21
KIRC	Kidney renal clear cell carcinoma	1437	376	132	158	32	19	109
LUAD	Lung adenocarcinoma	1780	363	114	169	38	15	117
LUSC	Lung squamous cell carcinoma	2096	402	111	158	49	16	96
THCA	Thyroid adenocarcinoma	888	207	65	103	27	7	72
TNBC	Triple-negative breast carcinoma	839	211	51	64	21	10	38
UCEC	Uterine corpus endometrioid carcinoma	1449	332	95	136	37	17	86

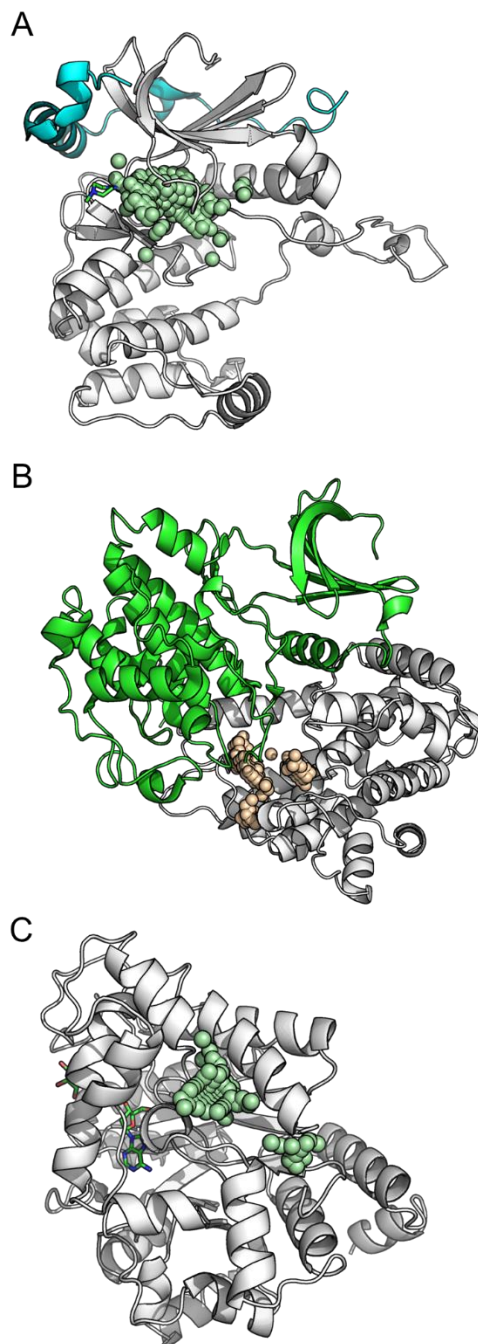


Figure 5.1. Examples of binding site annotations. Proteins are represented in cartoon format. The monomer structure with binding sites present is in white. SiteMap sites are shown as spheres, bound ligands are shown as sticks. **(A)** Enzyme (ENZ) site occupied by a bound inhibitor on the protein kinase domain of *AURKB* (PDB: 4AF3A). **(B)** PPI site at the interface of *CCNE1* (PDB: 1W98B) with *CDK2* (green). **(C)** OTH (Non-ENZ, non-PPI) site on *ADA* (PDB: 3IARA).

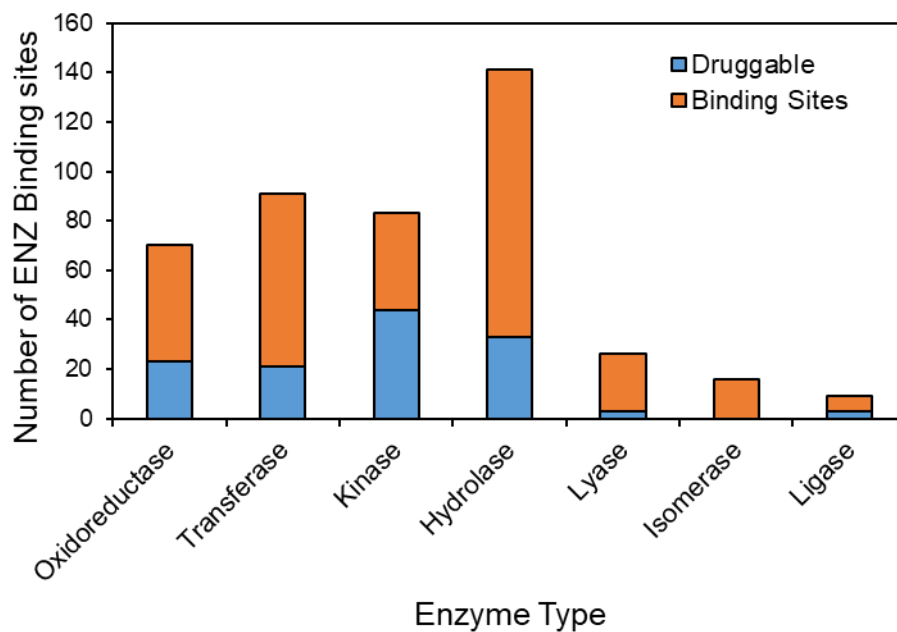


Figure 5.2. Classification of enzyme types by EC codes. Binding sites that were classified as enzyme (ENZ) through manual annotation via UniProt and Catalytic Site Atlas were classified using the protein’s EC codes. Binding sites were filtered using SiteScore and DrugScore greater than 0.8. Druggable binding sites feature a more stringent DrugScore cutoff of 1.0. While kinases are normally classified as part of the transferase family, here we have separated the two.

Among 1624 overexpressed proteins with at least one high-resolution human crystal structure, 1044 (~64%) had at least one binding site (**Table 5.1**). Similarly, among the 1218 highly overexpressed proteins with crystal structures, 405 (~33%) had at least one druggable binding site. For individual diseases, roughly 30% of proteins with crystal structures corresponding to highly overexpressed genes possessed at least one druggable binding site (**Table 5.2**). For example, 51 proteins with a crystal structure from among 211 in TNBC had a druggable binding site, while 114 proteins with a crystal structure in LUAD were found to have a binding site among 363. Generally, we found more binding sites than proteins with crystal structures, suggesting that although many of the proteins harbored more than one binding site, a large portion might only act as probes rather than druggable sites. An average of about 0.38 druggable binding sites were identified per protein with crystal structures. For example, a total of 145 druggable binding sites were identified on the 429 proteins with crystal structures corresponding to differentially expressed GBM genes. Among the most frequently overexpressed proteins with druggable binding sites are the members of the matrix metalloproteinases (MMPs) and protein kinases related to cell signaling.

5.2.3 Classification of Binding Sites. To characterize the potential functional impact of each of these binding sites, we classified each binding site by its functional role based on its structural features and location on the protein surface, particularly whether it corresponds to a catalytic site or to a binding site located at a protein-protein interaction interface. Using the proximity of known structural features and the functional annotations of key residues, we characterize each binding site on the protein structure of overexpressed genes from TCGA into three groups: enzyme (ENZ), protein-protein interaction (PPI), and other (OTH). **Fig. 5.1** shows examples of each of the three binding sites. For example, the ATP binding site of a protein kinase is classified as enzyme (ENZ), while a binding site at the interaction interface between two members of the protein families CDKs and cyclins are classified as PPI. All other binding sites are referred to as “other” (OTH). Within the binding sites that we identified, there is a wide distribution of binding site functions for each cancer type (**Table 5.1** and **5.2**). Overall, there are many more ‘OTH’ binding sites than ENZ and PPI across all tumors. OTH binding sites constitute approximately 70% of the binding sites observed, while ENZ and PPI are observed in about 20 and 10% of structures, respectively. Among those binding sites that we classify as druggable, the distributions are 25, 11, and 66% for the ENZ, PPI, and OTH binding sites, respectively. OTH binding sites may correspond to uncharacterized enzyme active sites or may occur at PPI interfaces that have not been characterized.

5.2.4 Cavities at Enzyme Active Sites. Enzyme active site binding sites were identified by first mapping known catalytic residues from Catalytic Site Atlas (CSA) [228] and UniProtKB

[227] onto the identified structures of each protein. CSA identifies catalytic residues as those that are (i) directly involved in a catalytic mechanism; (ii) alter the pK_A of another residue or water involved in the catalytic mechanism; (iii) stabilize a transition or intermediary state; and/or (iv) activate a substrate [228]. UniProt defines these residues as being directly involved in catalysis [227]. If one of the catalytic residues was within the binding site, we classify the binding site as ENZ. In total, we identified 434 unique enzyme active site binding sites and 126 druggable binding sites on proteins that are encoded by overexpressed genes at TCGA (**Table 5.1**). The number of druggable ENZ binding sites ranged from 10 for HNSC to 49 for LUSC. For example, there were 34, 21, and 38 druggable enzyme binding sites for GBM, TNBC and LUAD, respectively (**Table 5.2**). We further classify enzymes by their catalytic function and distinguish between the druggability of the binding site (**Fig. 5.2**). We treat kinases separately from the transferases. When kinases and transferases are combined, they, along with the hydrolases, are the largest group among the enzyme active site binding sites. There were 70, 91, 83, and 141 oxidoreductases, transferases, kinases, and hydrolases, respectively. Lyases, isomerases, and ligases, on the other hand, were the least common among proteins with ENZ binding sites (26, 16, and 9, respectively).

5.2.5 Cavities at Protein-Protein Interaction Interfaces. Despite the fact that protein-protein interactions play a crucial role in a range of diseases including cancer, few successful PPI inhibitors have been developed to date. This is attributed to the fact that PPI interfaces are usually large and devoid of well-defined binding cavities. Druggable binding sites that occur at protein-protein interfaces could be used to develop small molecules to disrupt the protein-protein interaction. PPI binding sites were identified by looking at the crystal structures with protein complexes with respect to the representative structures for a given protein. For each representative structure of a given protein, we went back to our sequence-based clustering approach in CD-HIT and identified the set of protein structures that shared significant sequence identity with the representative structure. We then aligned all the crystal structures from this alternative set of structures back onto the representative structure. This superimposition resulted in the identification of PPI interfaces that might not have appeared in the reference structure and their positions with respect to the previously identified binding sites. In total, we identified 231 unique binding sites located at protein-protein interaction interfaces, of which only 55 were druggable. As expected, there were significantly fewer binding sites that occurred at PPI interfaces than any of the other classes of binding sites. These ranged from 4 for HNSC to 19 for KIRC (**Table 5.2**).

5.2.6 Proteins with Binding Sites Located at Both Enzyme Active Sites and Protein-Protein Interaction Interfaces. While OTH binding sites were predominant among the different cancer types, the ENZ and PPI binding sites give greater insight into the binding site's function.

Interestingly, there are proteins that contain binding sites that are classified as both ENZ and PPI (**Table 5.3**). Of these 24 proteins, 10 have binding sites that are druggable and are part of the enzyme active site and a PPI interface. Among these are proteins that are implicated in cancer progression and metastasis, such as *CDA* [229] (**Fig. 5.3A**), *MMP14* [230] and *DDR1* [231]. In these cases, the binding site at the catalytic site is also part of a PPI interface. Many of the cases where the ENZ and PPI binding sites overlap correspond to binding sites that occur at the active site of proteases. The binding partner is usually a protease inhibitor, for example, *AGT* and *TIMPI* in *ANPEP* and *MMP14*, respectively. Generally, these interactions may not be promising targets since proteolytic activity may contribute to tumor invasion and metastasis. However, the overexpression of protease inhibitors such as TIMPs and serpins suggest that inhibition of proteases may oppose growth and metastasis of a tumor.

Other proteins contain distinct enzyme and PPI binding sites (**Table 5.4**). Of these 24 proteins, only *ALOX12* and *NR1L2* feature both druggable ENZ and PPI binding sites. These proteins can be placed into two categories based whether or not the binding sites are on the same protein domains. Some have ENZ and PPI binding sites on the same domain such as the decarboxylase *GADI*, which has a catalytic site as well as a PPI binding site at its homodimer interface. Another example is the phosphoribosyltransferase *NAMPT*, which is implicated in cancer metabolism [232], and has an ENZ binding site with an inhibitor bound as well as a PPI binding site between the homodimer structure (**Fig. 5.3B**). Other proteins have ENZ and PPI binding sites on separate domains. For example, the serine/threonine-protein kinase *PLK1* has both an enzymatic ATP binding site on its protein kinase domain and a binding site at the PPI interface at its POLO-box domain. Another similar example is the receptor tyrosine kinase *EPHB4*, which has an enzymatic ATP binding site on its protein kinase domain (**Fig. 5.3C**) and a binding site at the PPI interface with an ephrin ligand *EFNB2* on its ligand binding domain (**Fig. 5.3D**). These binding sites may be used to develop allosteric modulators. Small molecules that bind to the PPI binding site may alter substrate binding to the active site. A small-molecule inhibitor of enzyme activity may affect the protein-protein interaction of the protein.

Table 5.3. Proteins with binding site that is both ENZ and PPI.

		Interaction Partner		
		PDB	Symbol	Name
ANPEP	Aminopeptidase N	4FYSC	AGT	Angiotensinogen
CDA	Cytidine deaminase	1MQ0A	CDA	Cytidine deaminase
CTSV	Cathepsin L2	3KFQC†	CSTA	Cystatin-A
DDR1	Epithelial discoidin domain-containing receptor 1	3ZOSA	DDR1	Epithelial discoidin domain-containing receptor 1
DNM1	Dynamin-1	2X2ED	DNM1	Dynamin-1
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	1ZNQR†	GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GLA	Alpha-galactosidase A	3HG3B	GLA	Alpha-galactosidase A
GSG2	Serine/threonine-protein kinase haspin	4OUCB†	HIST2H3A	Histone H3.2
HDC	Histidine decarboxylase	4E1OE†	HDC	Histidine decarboxylase
HOGA1	4-hydroxy-2-oxoglutarate aldolase, mitochondrial	3SO5A†	HOGA1	4-hydroxy-2-oxoglutarate aldolase, mitochondrial
KIF3C	Kinesin-like protein KIF3C	3B6VB	KIF3C	Kinesin-like protein KIF3C
MMP14	Matrix metalloproteinase-14	3MA2B	TIMP1	Metalloproteinase inhibitor 1
PCSK9	Proprotein convertase subtilisin/kexin type 9	3BPSP†	PCSK9	Proprotein convertase subtilisin/kexin type 9
PGC	Gastricsin	1AVFQ	PGC	Gastricsin
PGD	6-phosphogluconate dehydrogenase, decarboxylating	2KJVC	PGD	6-phosphogluconate dehydrogenase, decarboxylating
PKLR	Pyruvate kinase PKLR	4IMAC	PKLR	Pyruvate kinase PKLR
PNLIPRP2	Pancreatic lipase-related protein 2	2PVSB†	PNLIPRP2	Pancreatic lipase-related protein 2
PNP	Purine nucleoside phosphorylase	4ECEE†	PNP	Purine nucleoside phosphorylase
REN	Renin	3G72A†	REN	Renin
RNASE2	Non-secretory ribonuclease	2BEXB	RNH1	Ribonuclease inhibitor
RRM1	Ribonucleoside-diphosphate reductase large subunit	2HNCB	RRM1	Ribonucleoside-diphosphate reductase large subunit
SEPT3	Neuronal-specific septin-3	3SOPB	SEPT3	Neuronal-specific septin-3
TDO2	Tryptophan 2,3-dioxygenase	4PW8E†	TDO2	Tryptophan 2,3-dioxygenase
UCHL1	Ubiquitin carboxyl-terminal hydrolase isozyme L1	3IFWB	UBC	Polyubiquitin-C

† The identified binding site is druggable (DS ≥ 1.0).

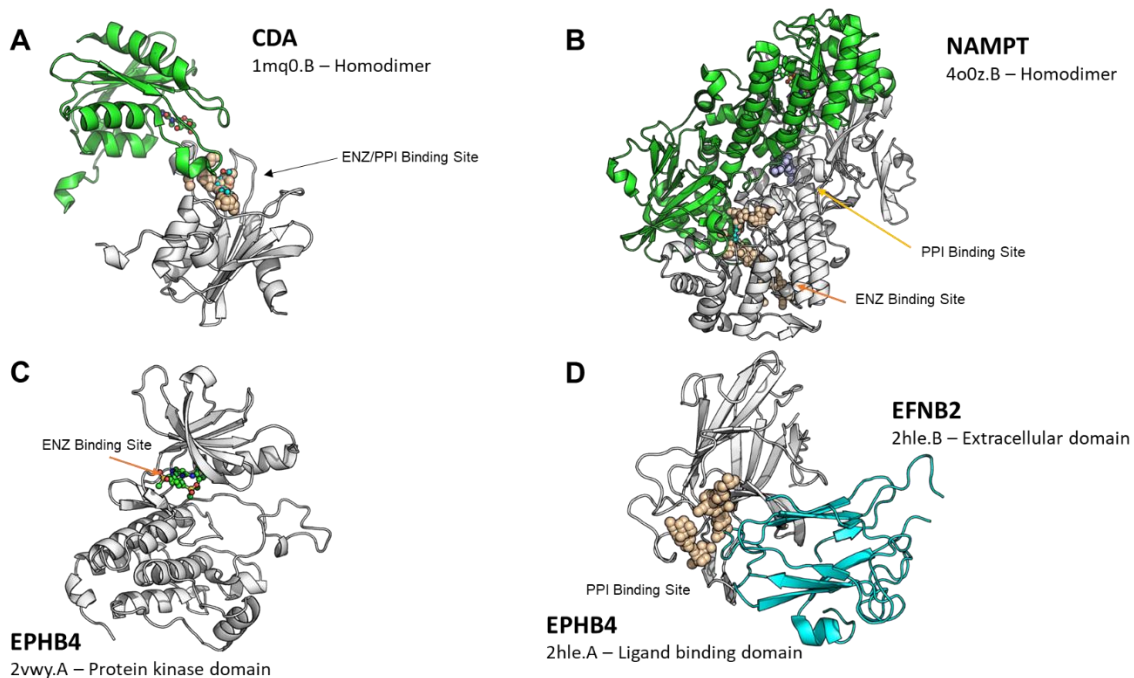


Figure 5.3. Examples of proteins with both ENZ and PPI binding sites. Proteins are represented in cartoon format. The monomer structure with identified binding sites is in white. SiteMap binding sites are shown as spheres, bound ligands are shown as ball-and-sticks. **(A)** The homodimeric structure of *CDA* (PDB: 1MQ0B) with a bound inhibitor at a binding site classified as both ENZ and PPI. **(B)** The homodimeric structure of *NAMPT* (PDB: 4O0ZB) with an ENZ (peach, bound inhibitor) and a PPI (blue) binding site on the same domain. **(C-D)** The protein kinase (PDB: 2VWYA) and ligand binding domain (PDB: 2HLEA) of *EPHB4* featuring an ENZ **(C)** and a PPI **(D)** binding site on separate domains. The binding site on the protein kinase domain is not shown as spheres but is occupied by the bound inhibitor (green).

Table 5.4. Proteins with both ENZ and PPI binding sites.

Symbol	Name	Interaction Partner		
		PDB	Symbol	Name
ACMSD	2-amino-3-carboxymuconate-6-semialdehyde decarboxylase	4IH3A	ACMSD	2-amino-3-carboxymuconate-6-semialdehyde decarboxylase
ADH1C	Alcohol dehydrogenase 1C	1HSOA	ADH1C	Alcohol dehydrogenase 1C
ALOX12	Arachidonate 12-lipoxygenase, 12S-type	3D3LB†	ALOX12	Arachidonate 12-lipoxygenase, 12S-type
AOC1	Amiloride-sensitive amine oxidase	3MPHB	AOC1	Amiloride-sensitive amine oxidase
BHMT	Betaine--homocysteine S-methyltransferase 1	1LT7B	BHMT	Betaine--homocysteine S-methyltransferase 1
CTSE	Cathepsin E	1TZSP	CTSE	Cathepsin E
DDC	Aromatic-L-amino-acid decarboxylase	3RFBF	DDC	Aromatic-L-amino-acid decarboxylase
DDX39A	ATP-dependent RNA helicase DDX39A	1T6NB	DDX39A	ATP-dependent RNA helicase DDX39A
EPHB2	Ephrin type-B receptor 2	2QBXD		Ephrin binding site
EPHB4	Ephrin type-B receptor 4	2HLEB	EFNB2	Ephrin-B2
GAD1	Glutamate decarboxylase 1	3VP6A	GAD1	Glutamate decarboxylase 1
GPI	Glucose-6-phosphate isomerase	1JIQB	GPI	Glucose-6-phosphate isomerase
HK2	Hexokinase-2	2NZTA	HK2	Hexokinase-2
HMGCS2	Hydroxymethylglutaryl-CoA synthase, mitochondrial	2WYAD	HMGCS2	Hydroxymethylglutaryl-CoA synthase, mitochondrial
NAMPT	Nicotinamide phosphoribosyltransferase	4O0ZA	NAMPT	Nicotinamide phosphoribosyltransferase
NR1I2	Nuclear receptor subfamily 1 group I member 2	3CTBB†	NR1I2	Nuclear receptor subfamily 1 group I member 2
NTRK1	High affinity nerve growth factor receptor	1WWWV	NGF	Beta-nerve growth factor
PLK1	Serine/threonine-protein kinase PLK1	1Q4KE		Phosphopeptide
PYGL	Glycogen phosphorylase, liver form	2ZB2B	PYGL	Glycogen phosphorylase, liver form
RHOC	Rho-related GTP-binding protein RhoC	3KZ1A	ARHGEF11	Rho guanine nucleotide exchange factor 11
SULT1C2	Sulfotransferase 1C2	3BFXA	SULT1C2	Sulfotransferase 1C2
TH	Tyrosine 3-monooxygenase	2XSNC	TH	Tyrosine 3-monooxygenase
TPH2	Tryptophan 5-hydroxylase 2	4VO6B	TPH2	Tryptophan 5-hydroxylase 2
UPP1	Uridine phosphorylase 1	3EUFB	UPP1	Uridine phosphorylase 1

† The identified binding site is druggable (DS ≥ 1.0).

5.2.7 Unclassified Binding Sites. Binding sites that were neither enzyme active sites nor located at protein-protein interactions were classified as OTH. In total, more than 1500 of these binding sites were identified on proteins that are encoded by differentially expressed genes. These binding sites could potentially be either unassigned enzyme active sites, part of structurally unresolved protein-protein interaction sites, or allosteric sites. A binding site is considered allosteric only if it occurs on a protein that has enzyme activity or that engages other ligands at sites that are distant from the allosteric binding site. Among the 782 proteins with OTH binding sites, 323 also have at least one ENZ or PPI binding site. These binding sites offer an opportunity to design allosteric small molecule modulators of enzyme activity or protein-protein interactions. Allosteric regulation of enzyme activity has been successfully achieved with small molecules in several systems [233]. For example, small molecule kinase inhibitors have been developed to bind to allosteric binding sites to inhibit the enzyme activity of the protein kinase [234]. More recently, small molecules that bind to an allosteric binding site on the Ral GTPase was shown to modulate the distal interaction with its effector protein [235].

Many OTH binding sites occur on proteins with existing ENZ and/or PPI binding sites, which may be potential allosteric sites for protein inhibition. When the enzyme active site is well characterized on a protein surface, additional binding sites represent opportunities for allosteric inhibition of the protein's function. For example, the sulfotransferase *SULT2B1* has four binding sites on its protein surface (**Fig. 5.4A**). The ENZ binding site is encompassed by the adenosine nucleotide. Three additional OTH binding sites were detected on the surface of the protein and represent potential sites for allosteric sites. Another example of protein with both ENZ and OTH binding sites is the protein kinase *RET* (**Fig. 5.4B**). In this structure, a known inhibitor occupies the ENZ ATP binding site, while an additional allosteric binding site is formed near the α C helix. Similarly, there are proteins with both PPI and OTH binding sites. One example is the PPI between *CHN2* and *SLC9A1* (**Fig. 5.4C**), where an α -helix from *SLC9A1* occupies two PPI binding sites on *CHN2*. An additional potentially allosteric OTH binding site is formed on the backside of *CHN2*. Another example is the protein complex formed between *PLAUR*, *PLAU*, and *VTN* (**Fig. 5.4D**). In this example, binding sites were found on the monomer structure of the apo protein. After superimposition of additional crystal structures back onto the representative structure, two of the three detected binding sites were classified as PPI. The two separate PPI binding sites occupy the respective interfaces between *PLAUR-PLAU* and *PLAUR-VTN*. An additional OTH binding site was also detected on the protein surface and represents an allosteric site.

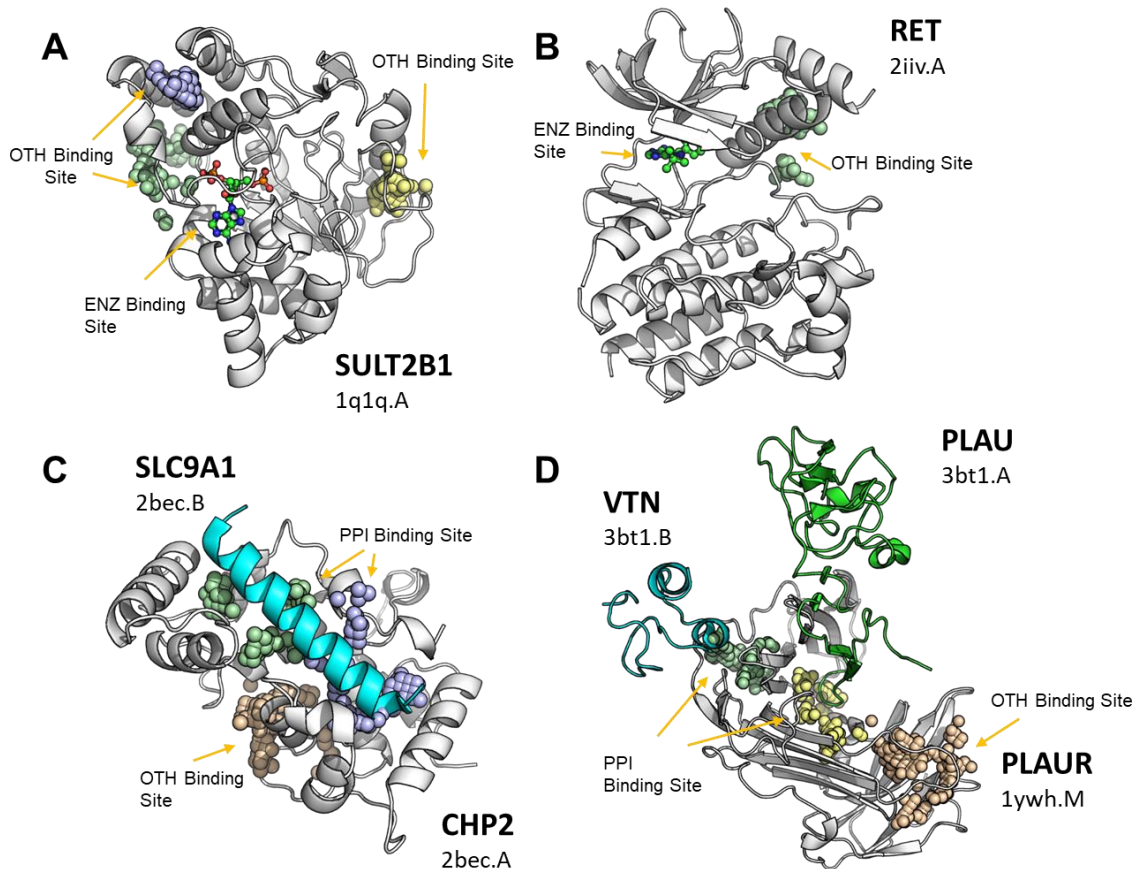


Figure 5.4. Examples of proteins with potentially allosteric OTH binding sites. Proteins are represented in cartoon format. The monomer structure with identified binding sites is in white. SiteMap binding sites are shown as spheres, bound ligands are shown as ball-and-sticks. **(A)** *SULT2B1* (PDB: 1Q1QA) with an ENZ binding site occupied by a nucleotide and three additional OTH binding sites (green, blue, yellow). **(B)** *RET* (PDB: 2IIVA) with an ENZ binding site occupied by the bound inhibitor and an additional OTH binding site (green). **(C)** *CHP2* (PDB: 2BECA) with two PPI binding sites (green, blue) at the interface with *SL9CA1* (PDB: 2BECB) and an additional OTH binding site (peach). **(D)** The superimposed structure of *PLAUR* (PDB: 1YWHM) with two PPI binding sites at the interfaces with *VTN* (PDB: 3BT1B, green) and *PLAU* (PDB: 3BT1A, yellow) and an additional OTH binding site (peach).

Table 5.5. Proteins with potential PPI binding sites identified from search against PrePPI.

Symbol	Name	Binding site	Predicted PPI		
			Model	Symbol	Name
AK3	GTP:AMP phosphotransferase AK3, mitochondrial	1ZD8A2	2BWJ	AK5	Adenylate kinase isoenzyme 5
ANK1	Ankyrin-1	1N11A3	2JAB	ILK	Integrin-linked protein kinase
CHN1	N-chimaerin	3CXLA3	1OW3	RAC1	Ras-related C3 botulinum toxin substrate 1
HOGA1	4-hydroxy-2-oxoglutarate aldolase, mitochondrial	3S5OA1†	3DAQ	HOGA1	4-hydroxy-2-oxoglutarate aldolase, mitochondrial
HPD	4-hydroxyphenylpyruvate dioxygenase	3ISQA1†	1SQI	HPDL	4-hydroxyphenylpyruvate dioxygenase-like protein
LCN	Lipocalin-1	3EYCA1†	2F91	OVCH1	Ovochymase-1
NCS1	Neuronal calcium sensor 1	1G8IB2†	1AUI	PPP3CA	Serine/threonine-protein phosphatase 2B catalytic subunit alpha isoform
RAP1GAP	Rap1 GTPase-activating protein 1	1SRQA1	3BRW	RAP1A	Ras-related protein Rap-1A
RHCG	Ammonium transporter Rh type C	3HD6A1	2NUU	RHAG	Ammonium transporter Rh type A
SHMT2	Serine hydroxymethyltransferase, mitochondrial	3OU5A1	3GBX	SHMT2	Serine hydroxymethyltransferase, mitochondrial
STXBP2	Syntaxin-binding protein 2	4CCAA2	3C98	STX1A	Syntaxin-1A
THEM5	Acyl-coenzyme A thioesterase THEM5	4AE7A1	1Q4T	THEM4	Acyl-coenzyme A thioesterase THEM4
ZBTB32	Zinc finger and BTB domain-containing protein 32	3M5BB1	3BIM	BCL6	B-cell lymphoma 6 protein

† The binding site is druggable (DS \geq 1.0)

5.2.8 A Search of Protein-Protein Interaction Networks to Identify OTH Binding Sites

Located at PPI Interfaces. The majority of OTH binding sites occur on proteins with no discernable ENZ or PPI binding sites. To determine whether these binding sites could potentially be located at protein-protein interaction interfaces, a database of predicted protein-protein complexes known as PrePPI was explored [236]. The PrePPI method uses both structural and non-structural evidence to predict whether two proteins form a complex. For complexes predicted based on structural information, PrePPI superimposes monomeric crystal structures onto a reference complex based on the structural similarities of the monomeric structures with the two structures forming the interaction interface. This model is then evaluated based on how well the individual residues of the predicted interaction interface overlap with the structural model. If the likelihood ratio of this structural modeling is above a given cutoff, PrePPI provides the identifiers of both the individual proteins and the reference structure for further evaluation. For the 458 proteins that contained only binding sites classified as OTH, we evaluated the structural models given by PrePPI to determine whether or not OTH binding sites overlapped with potential PPI interfaces. These 458 proteins are represented by 395 unique crystal structures consisting of 806 binding sites of unknown function. Of these 806 OTH binding sites, 48 were on proteins without models of structural complexes in PrePPI. Among the remaining 758 OTH binding sites, we identified 17 OTH binding sites on 13 proteins that are likely binding sites at protein-protein interfaces (**Table 5.5**). In each of these 17 cases, a previously classified OTH binding site was predicted by PrePPI to be part of a known protein-protein interaction interface, and perhaps directly contributing to the PPI itself. It is interesting to note that several of these predicted protein-protein interactions are well-established despite the lack of a co-crystal structure: These include the *ANK1-ILK* [237] and *CHN1-RAC1* [238] interactions. In each of these cases, there was high degree of homology between the structure containing the OTH binding site and the PrePPI protein-protein complex to which it was superimposed. In most cases, however, the protein containing the OTH binding site did not show any homology with a protein in a PrePPI complex. In these cases, the similarity between the interaction interfaces of the two proteins and a model protein complex was used. The *NCSI-PPP3CA*, *LCN1-OVCH1*, and *ZBTB32-BCL6* interactions are examples in which the interaction was uncharacterized in both the literature and existing PPI databases. These three interactions were predicted based on the structural complementarity of both the interaction interface and the crystal structure. Overall, we predict that approximately 2% of OTH binding sites with unknown function to be part of a previously uncharacterized PPI interface.

5.2.9 Cancer Signaling Pathways. Pathways reveal signaling transduction across a cascade of proteins that elicit a variety of cell phenotypes. Individual targets in these pathways are

potential sites through which small-molecule inhibition are expected to enhance or alter the subsequent cell phenotype. Alteration of individual genes within these signaling pathways lead to cancer related processes such as cell growth and adhesion. We have identified 27 cancer related signaling pathways in KEGG [239] and their respective proteins. Using the members in each of these signaling pathways, we map binding sites onto these individual proteins. We distinguish between binding sites with DrugScore greater than 0.8 on proteins with \log_2 fold change greater than 1.5 (i.e., able to be probed) (**Fig. 5.5A**) and those with DrugScore greater than 1.0 and \log_2 fold change greater than 2 (i.e., druggable binding sites) (**Fig. 5.5B**). While some signaling pathways like the cell cycle contained binding sites of all functional types, no binding sites could be identified for the Hedgehog pathway on differentially expressed genes.

To address crosstalk between signaling pathways, binding sites were also evaluated as being either unique to that signaling pathway or on proteins that occur in multiple signaling pathways. In a majority of cancer signaling pathways, there were more binding sites that occurred in multiple signaling pathways than in a signaling pathway, revealing proteins targets that are involved in multiple signaling processes. Only the Citrate Cycle, HIF-1, and PPAR signaling pathways had many more binding sites that were unique to the signaling pathway itself than in multiple signaling pathways. In signaling pathways such as focal adhesion and cytokine-cytokine receptor interactions, almost all of the druggable binding sites belonged to proteins that were involved in crosstalk across cancer signaling pathways. Finally, signaling pathways such as the cell cycle and Hippo pathways have an even mix of binding sites on unique and overlapping proteins.

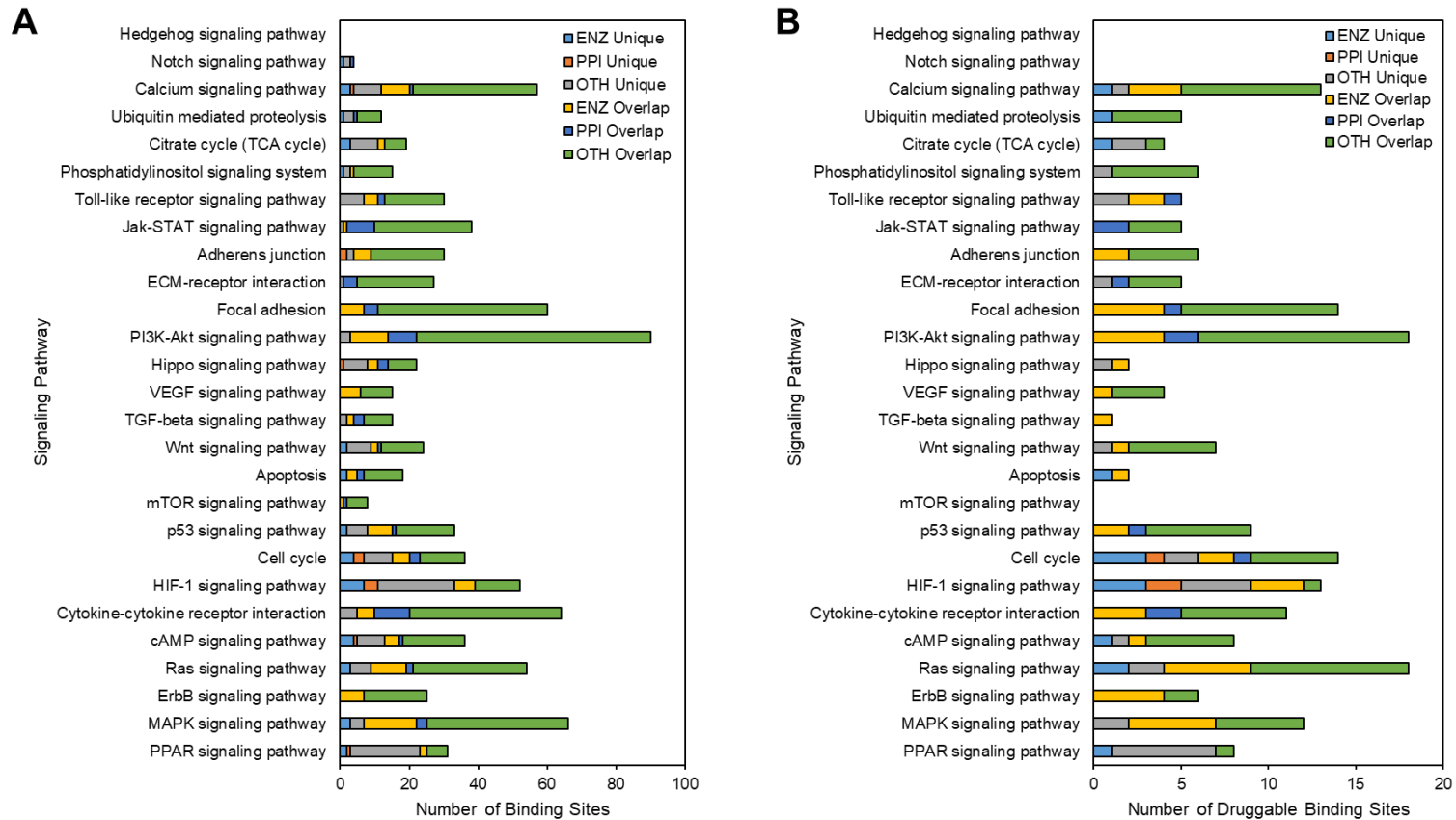


Figure 5.5. Binding sites in cancer-related signaling pathways. Proteins with binding sites were mapped to 27 cancer related signaling pathways in KEGG. Identified binding sites were divided based on whether the protein was exclusive to one signaling pathway or occurred in multiple signaling pathways. **(A)** Identified binding sites had DrugScore greater than 0.8 on proteins with \log_2 fold change greater than 1.5. **(B)** Identified binding sites had DrugScore greater than 1.0 and \log_2 fold change greater than 2.

5.2.10 Correlation with Patient Survival for Proteins Encoded by Differentially Expressed Genes. We collected patient survival data from TCGA clinical records for each disease to identify the impact of gene expression on overall survival of cancer patients. To determine the overall survival rate, we first identified the date of death or date of the last checkup for deceased and living patients, respectively. For each differentially expressed gene among the 10 diseases we considered, the median expression value was used to divide patient tumors into two groups, high and low expression. For a given gene, we then paired a patient's gene expression with their survival outcome to build a Cox proportional hazards regression model for differentially expressed genes. The ratio of the hazard rates between the high and low expression groups are summarized by a metric known as the hazard ratio. The hazard ratio derived from the regression model defines the probability that an event will occur in the next time interval. In this model, this time interval is made sufficiently small that the hazard rate is considered instantaneous. Therefore, the hazard ratio is used to describe the ratio between the hazard rate of two groups, that is, the survival of patients expressing a gene at high and low levels.

In total, we identified 1343 differentially expressed genes across all 10 diseases with a hazard ratio above 1 and \log_2 fold change above 1.5. Among them, 202 contained at least one binding site (**Fig. 5.6A**). Both KIRC (121 total) and LUAD (57 total) had the greatest number of proteins that were both overexpressed and correlated with patient outcome. There were 45 druggable genes that were found to be both overexpressed and correlated with patient outcome in more than one cancer type. The most frequently occurring are *MELK* and *RRM2* in 4 separate cancers, while another 9 proteins have significant fold changes and hazard ratios in 3 cancers. The binding sites on these 202 proteins show a wide distribution in both their druggability and binding site type (**Fig. 5.6B**).

Of the 601 unique binding sites on these proteins, 102 are ENZ, 46 are PPI, 444 are OTH, and 9 are classified as both ENZ and PPI. Both the SiteScore and DrugScore of the PPI binding sites have upper limits of about 1.1 for both metrics, while there are many ENZ and OTH binding sites that exceed this cutoff. Similarly, we focused on the subset of the proteins that were highly overexpressed and featured druggable binding sites. In total, we identified 60 proteins with at least one druggable binding site across 10 diseases with a \log_2 fold change greater than 2.0 and hazard ratio greater than 1.0 (**Fig. 5.6C**). Similarly, there are far fewer binding sites among proteins that fit these criteria. Of the 92 binding sites, 20 are ENZ, 6 are PPI, 65 are OTH, and 1 is both ENZ and PPI (**Fig. 5.6D**).

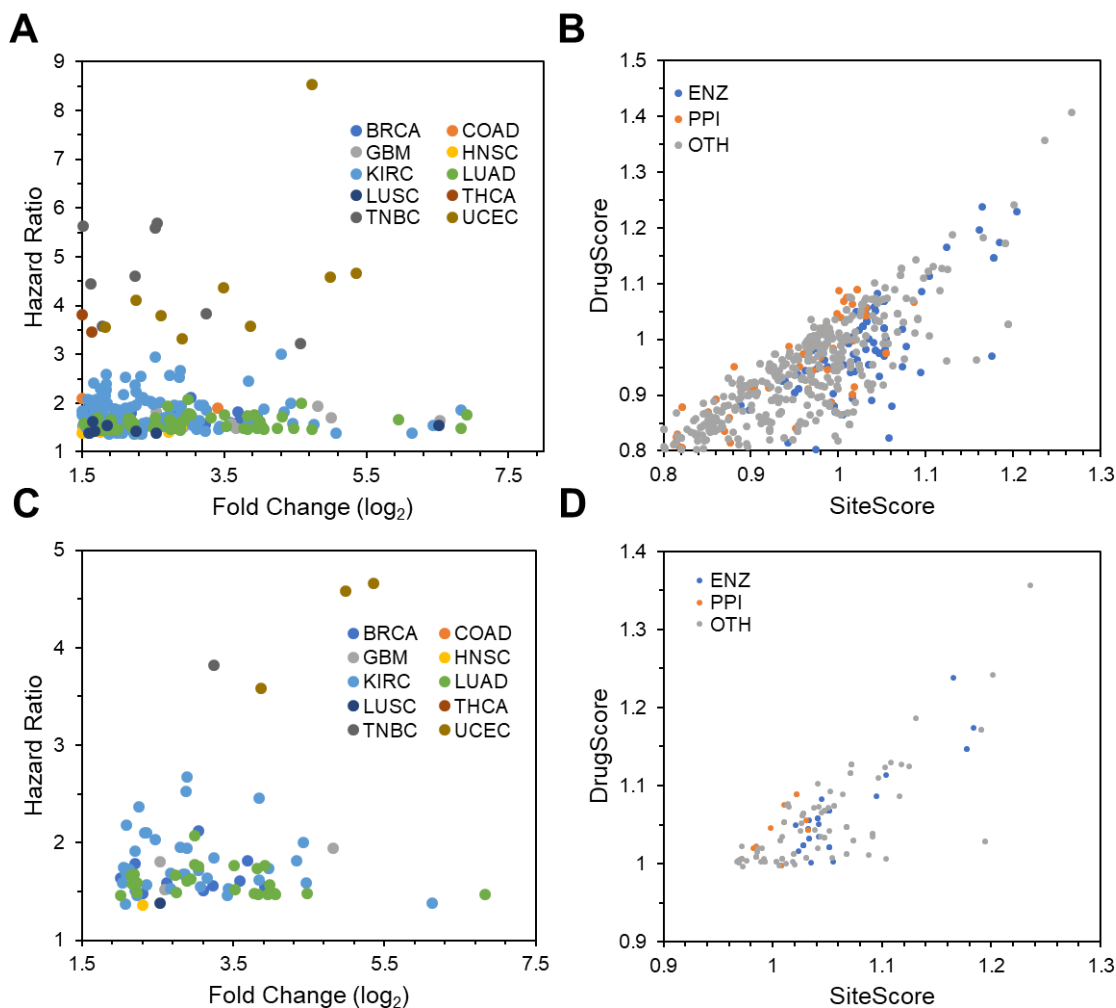


Figure 5.6. Proteins with binding sites that are both overexpressed and correlate with patient outcome. **(A)** Fold change versus hazard ratio across all cancer types on proteins with $\log_2FC \geq 1.5$, $HR > 1.0$, and $DrugScore > 0.8$. **(B)** SiteScore and DrugScore of binding sites by functional annotation. **(C)** Fold change versus hazard ratio across all cancer types on proteins with druggable binding sites with $\log_2FC \geq 2.0$, $HR > 1.0$, and $DrugScore > 1.0$. **(D)** SiteScore versus DrugScore of druggable binding sites with $\log_2FC \geq 2.0$, $HR > 1.0$, and $DrugScore > 1.0$.

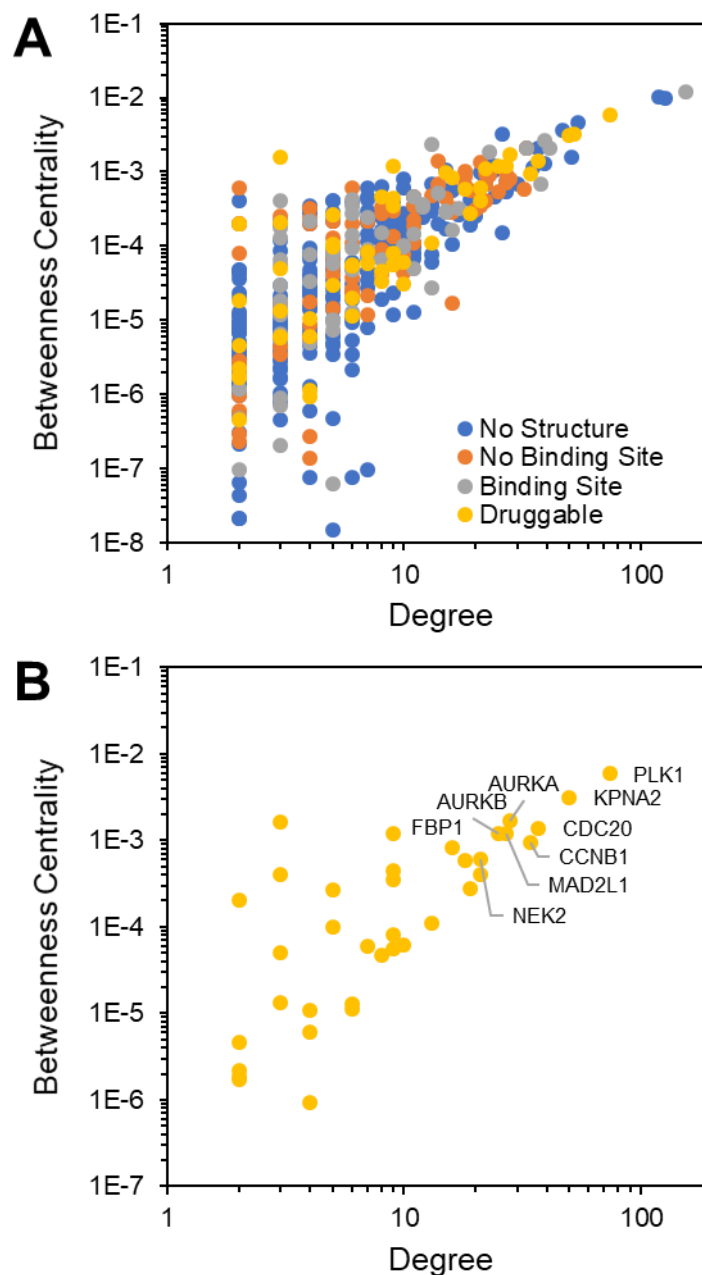


Figure 5.7. Integrating druggable binding sites with protein-protein interaction networks. **(A)** Degree versus betweenness centrality from PPI network for all proteins with $\log_2FC \geq 1.5$ and $HR > 1$. Proteins are color coded based on whether there was a high-quality crystal structure (blue), a crystal structure but no identifiable binding sites (orange), binding sites with DrugScore between 0.8 and 1.0 (gray), and druggable binding site with DrugScore greater than 1.0 (yellow). **(B)** Degree versus betweenness centrality from PPI network for all proteins with $\log_2FC \geq 2.0$, $HR > 1.0$, and $DrugScore > 1.0$.

5.2.11 Protein-Protein Interaction Network. In addition to looking at differentially expressed genes in the context of their expression, we addressed their impact on the global protein-protein interaction network. Networks have been used to not only model biological relationships, such as the relationship between drugs and diseases [240] or genes and diseases [241], to understand their underlying mechanisms, but also to identify new drug targets by identifying the relationships between a drug's side effects [242] or gene expression profile [243]. Using experimental data, a global protein-protein interaction network was constructed from physical interactions in humans by integrating data from seven major interaction databases. This resulted in 203068 non-redundant protein-protein interactions. To address the robustness of the network, we further filtered the interactions by only keeping those interactions that appeared in at least two of the seven databases. This resulted in a network with 38164 non-redundant protein-protein interactions.

We then identified the network properties of each protein within this network to measure the centrality and essentiality of each protein to the overall network. Among the topological properties of a given protein are its degree, which describes the number of interactions that are formed by that protein, and its betweenness centrality, which describes the number of shortest paths that go through the given protein. In a biological context, betweenness centrality is a measure of the available paths that a signal can travel through a given network [244]. Thus, proteins with high betweenness are thought to be essential to biological function and are frequently targeted in drug discovery [245]. For example, *TP53* has a betweenness centrality and degree of 4.1×10^{-2} and 236, respectively, while *EGFR* is 2.3×10^{-2} and 181 for the same properties. We examine the topological properties of all proteins that are overexpressed ($\log_2 \text{FC} \geq 1.5$) and whose expression correlate with patient outcome (**Fig. 5.7A**). Of these 1343 proteins, 1001 (~75%) did not have a high-quality crystal structure and an additional 141 (~10%) had a structure but no binding sites. Of the remaining proteins, 117 (9%) and 84 (6%) have binding sites and druggable binding sites, respectively. When the differential-expression cutoff is increased to 2 and the minimum DrugScore is increased to 1.0, 60 proteins have at least one druggable binding site (**Fig. 5.7B**). Among the proteins with the highest centrality and degree are *PLK1*, *KPNA2*, *AURKA*, and *AURKB*.

5.2.12 New Unexplored Targets for the Development of Small-Molecule Probes and Cancer Therapeutics. For each of the previously identified 60 targets, we integrate their structural, genomic, biological, and clinical data to examine their druggability. We divide these targets into those that are already established in cancer and those that are uncommon or novel based on the number of citations found in PubMed. Similarly, we analyzed the 202 proteins that were identified using the lower cutoffs in fold change and binding site DrugScore. We rank-ordered the top targets

for each cancer based on their interconnectivity in the PPI network. Among these potential targets, we see a variety of biological processes represented, including many involved in the immune response, metabolism, homeostasis and cell cycle. Similarly, some are well-studied in cancer but lack small-molecule inhibitors, while others have no co-crystallized small-molecule inhibitors but inhibitors have been reported in the literature. For example, the well-studied transcription regulator *TOP2A* is altered in cancer cells resulting in chromosome instability and is among the genes that are overexpressed and correlate with survival, but has many topoisomerase-specific inhibitors available [246]. Other genes may act as markers for cancer and indicate late progression into cancer or are vital to the immune response against tumorigenesis. However, there are many targets whose biology and lack of potential inhibitors may prove to be interesting targets for future considerations.

We identify examples of proteins with ENZ binding sites that have seldom been considered in cancer and lack therapeutics (e.g. *PYCR1*, *QPRT*, *HSPA6*), or are well-studied in cancer but lack small-molecule inhibitors (e.g. *PKMYT1*, *STEAP3*, *NNMT*). Similarly, we highlight examples of proteins with PPI binding sites that have not been previously targeted by small-molecule inhibitors and are either seldom considered in cancer (e.g. *CASC5*, *ZBTB32*, and *CSAD*), or are well-studied in cancer but lack small-molecule inhibitors (e.g. *HNF4A*, *MEF2B*, and *CBX2*). OTH binding sites can provide an avenue to modulate either enzymatic function or protein-protein interactions of the target. Compounds that bind to OTH sites could act either in an orthosteric manner if the binding site happens to be the binding site of a substrate or protein, or allosterically if the binding site is outside an enzyme active site or protein binding site. Among the genes whose overexpression strongly correlated with patient outcome and that possessed an OTH binding site, several had never been studied in cancer before nor do they have small-molecule inhibitors either in the literature or in co-crystallized complexes. Among these are four examples that span a variety of tumors: a protein of unknown function *FAM83A*, a water channel *AQP2*, a serine protease *SERPIND1*, and a protein associated with the immune response *TNFAIP8L2*.

Among these targets, 26 have been previously probed with small-molecule ligands and X-ray crystallography. Interestingly, many of these co-crystallized structures occur at binding sites at or below our higher DrugScore cutoff of 1.0, suggesting that a more stringent cutoff may discard otherwise druggable binding sites. Additionally, we mapped these druggable binding sites to conserved protein domains and find that these binding sites are mainly parts of the protein kinase, serpin, kinesin, and peptidase domains. When we consider only those without co-crystallized small-molecule inhibitors, protein kinases and trypsin domains are removed. The majority of binding sites across both targeted and untargeted proteins are classified as OTH. In well-studied systems where the active site is known, these OTH sites represent opportunities for allosteric regulation.

We next looked at the secondary structure of residues that compose the individual binding sites of these proteins across their individual binding site annotations. By examining the residues around a binding site, we generalized the type of secondary structures that were used to construct the binding site itself (**Fig. 5.8**). The majority of binding sites identified were a mixture of secondary structures or random coils among all proteins with or without small-molecule inhibitors. Combined, these two secondary structures generally making up the large majority of all binding sites in each binding site type. In each case, the least frequently observed secondary structure among these binding sites were the helix-like (i.e. α -helix, 3_{10} helix, or π -helix) and sheet-like structures (i.e. beta bridges and beta bulges). We then examined the secondary structures of the residues of the binding partner inside PPI binding sites. About 27 and 46% of the residues of the binding partners in the binding site were coil-like and helical (α -helix, 3_{10} helix, or π -helix), respectively. Only 10% of the binding sites were characterized by strand-like structures (β -sheet or β -bridge). The remaining PPI binding sites were a combination of these.

5.2.13 Missense Mutations on Protein Structures. A set of somatic mutations were obtained from a recent study from TCGA's Pan-Cancer initiative [247]. We identified missense mutations from this study onto patients in 7 of 10 diseases and mapped these to protein structures. We classified these mutations as being (i) adjacent to a binding site; (ii) elsewhere on the protein surface; or (iii) buried in the interior of the protein (**Fig. 5.9A**). We find that the majority of these missense mutations are found on the surface of proteins but not within a predicted binding site. The frequency of mutations occurring in the interior of a protein is higher than the frequency of mutations that occur at binding sites. We explored some of the proteins with mutations occurring most frequently in the binding site (**Fig. 5.9B**). They include well-known genes that have been previously reported to be heavily mutated in cancer such as *PIK3CA* [248], *SI* [249], and *PTEN* [250]. On the most commonly mutated target, *PIK3CA*, mutation rates are approximately five-fold less at the binding site than the entire protein. Also, among the top targets is *BRAF*, which features the common V600E mutation, and has been used for the rational design of small-molecule inhibitors of the mutant protein [251-253].

We matched these proteins with missense mutations with their gene expression levels and correlation with patient outcome. We find 29 binding sites on 26 proteins that are (i) overexpressed (\log_2 fold change ≥ 2); (ii) correlate with patient outcome (hazard ratio > 1); and (iii) have a missense mutation adjacent to a binding site in a given disease (**Table 5.6**). These 29 binding sites include 9 ENZ, 3 PPI, and 17 OTH pockets. Among these mutations adjacent to binding sites is the W167L mutation on the PPI interface between *MAD2L1* and *MAD1L1* in LUAD (**Fig. 5.9C**). This interaction is part of the spindle assembly checkpoint in the cell cycle [254]. Considering the

significant reduction in contact area upon replacing tryptophan with leucine, and the fact that tryptophan residues tend to often occur at protein-protein interaction interfaces, we expect that this mutation may impair the protein-protein interaction. Another mutation is the R121P mutation adjacent to the DNA-binding OTH binding site on *EXO1* in LUAD (**Fig. 5.9D**). The DNA-binding protein is also involved in DNA repair during cell cycle regulation [255]. Unlike the previous mutation, arginine contains a positively charged group while proline is a neutral non-polar amino acid.

We examined the mutation rates of individual amino acids by looking at the wild-type and mutated amino acids as a result of a mutation at each of the three locations on the protein (**Fig. 5.10**). We find differences in the relative frequencies of specific point mutations between each location. For example, mutations to alanine is less favored in the pocket or on the surface of the protein than it is in the interior, especially at charged or polar groups. Among the most common mutations in the binding site and on the surface is from lysine to glutamic acid, which occurs at a much lower frequency in the interior of the protein.

5.3. DISCUSSION

The sequencing of the genome of human tumors has provided access to an unprecedented number of new opportunities for the development of cancer therapeutics. While biological methods such as siRNA or CRIPSR/Cas9 methods are useful tools to explore the role of potential targets, chemical tools provide a complementary approach to interrogate new targets. Small molecules do not affect the expression of the target thereby causing little disruption to the signaling networks. In addition, small molecules have significantly greater precision as they can be designed to binding to a single cavity within a protein and modulate the function of the protein by disruption of protein-protein interactions or enzyme activity. Small molecules can work either in an orthosteric manner if they directly interfere with the binding of a protein or a substrate. They can also work in an allosteric manner by binding to cavities located outside protein-protein and protein-substrate binding interfaces and modulating the conformation and dynamics of the target.

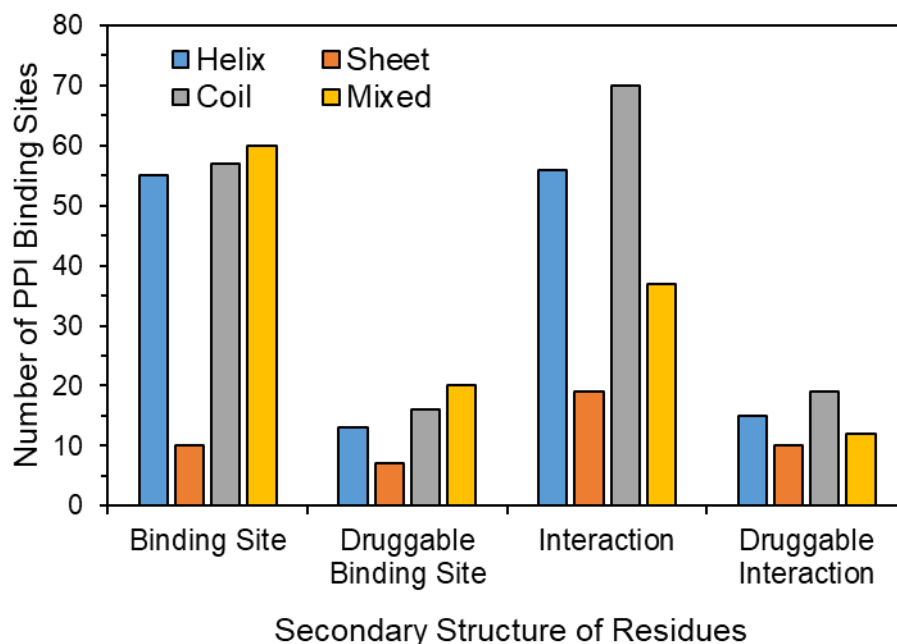


Figure 5.8. Secondary structure composition of residues surrounding PPI binding sites PPI binding sites. The secondary structure composition of both the binding site and the binding partner within the binding site was identified by creating a 5 Å sphere around the center of each binding site. Secondary structures were obtained from DSSP and combined based on whether the residues were primarily helix-like (i.e. α -helix, 3_{10} helix, or π -helix), sheet-like (i.e. beta bridges and beta bulges), random coils, or a mixture of these types.

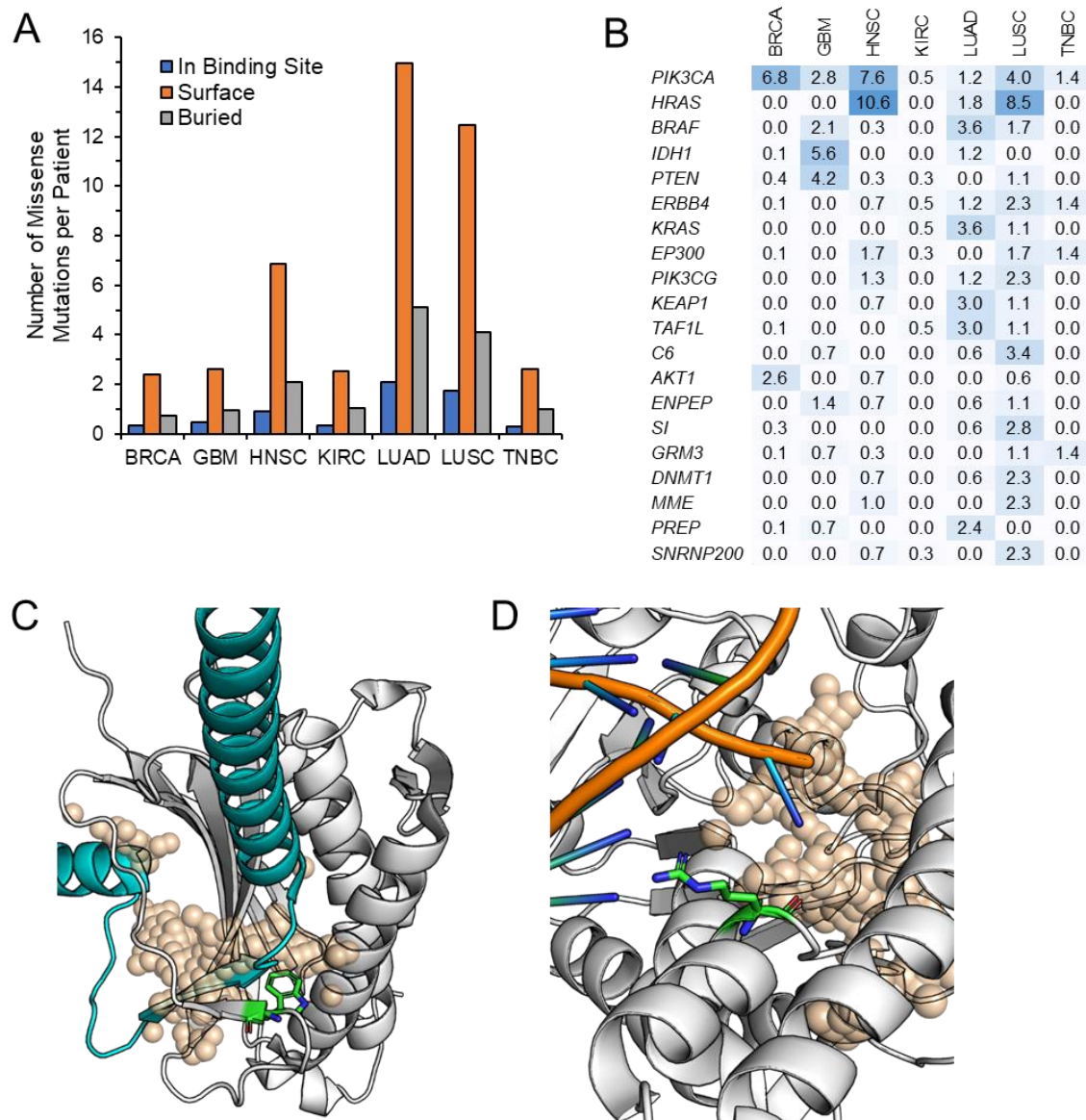


Figure 5.9. Proteins with missense mutations. **(A)** Missense mutations were mapped to patients in 7 of 10 diseases (COAD, THCA, and UCEC not included). Individual mutations were mapped to the protein structure and classified as being adjacent to the binding site, elsewhere on the protein surface, or buried in the interior of the protein structure. **(B)** Percentage of samples with missense mutations adjacent to a binding site in a given disease, showing the top 20 proteins rank-order using the sum of frequencies. **(C)** The W167L (green stick) mutation on the PPI interface between *MAD2L1* (white) and *MAD1L1* (cyan) is shown in cartoon (PDB ID: 1GO4). The PPI binding site is shown as transparent spheres. **(D)** The R121P (green stick) mutation adjacent to the DNA-binding OTH site (tan, transparent spheres) on *EXO1* (white cartoon) (PDB ID: 3QEB). DNA in the binding site from the crystal structure is also shown as cartoon.

Table 5.6. Mutations in binding site on overexpressed and clinically-relevant genes.

Symbol	Name	Cancer		Pocket	Type
		Type	Mutation		
ADH1C	Alcohol dehydrogenase 1C	LUAD	G205C	1HSZA1	ENZ
ADORA2A	Adenosine receptor A2a	BRCA	R293P	3VG9A5	PPI
C3	Complement C3	KIRC	C873Y	2WIIB4	OTH
CA6	Carbonic anhydrase 6	LUSC	H113Q	3FE4A1	ENZ
CCNA2	Cyclin-A2	LUAD	L341F	2BPMD1	OTH
CCNE1	G1/S-specific cyclin-E1	BRCA	A338T	1W98B2	OTH
CHEK1	Serine/threonine-protein kinase Chk1	LUAD	V46A	2R0UA1	ENZ
CYP2A6	Cytochrome P450 2A6	LUAD	V306I	2PG6B1	OTH
CYP2D6	Cytochrome P450 2D6	KIRC	L213P	3QM4A1	OTH
EXO1	Exonuclease 1	LUAD	R121P	3QEBZ1	OTH
F2	Prothrombin	KIRC	R543L	4NZQA3	OTH
KIF15	Kinesin-like protein KIF15	LUSC	G41A	4BN2C2	OTH
KIFC1	Kinesin-like protein KIFC1	LUAD	G568W	2REPA1	ENZ
MAD2L1	Mitotic spindle assembly checkpoint protein MAD2A	LUAD	W167L	2V64F1	PPI
MELK	Maternal embryonic leucine zipper kinase	BRCA	Q115R	4UMUA2	OTH
		LUAD	V271A	4UMUA2	OTH
NEK2	Serine/threonine-protein kinase Nek2	LUAD	R140L	2XK4A1	ENZ
PCK1	Phosphoenolpyruvate carboxykinase, cytosolic [GTP]	LUAD	R137H	2GMVA3	OTH
			A287S	2GMVA1	ENZ
			G289W		
PSPH	Phosphoserine phosphatase	LUSC	M52T	1L8OA1	ENZ
RHCG	Ammonium transporter Rh type C	LUAD	Q107H	3HD6A1	PPI
RRM2	Ribonucleoside-diphosphate reductase subunit M2	LUAD	E207Q	2UW2A2	OTH
SERPINB3	Serpin B3	LUAD	A45T	2ZV6A3	OTH
SERPINB4	Serpin B4	LUAD	S33N	2ZV6A2	OTH
SULT4A1	Sulfotransferase 4A1	KIRC	M80R	1ZD1A1	ENZ
TOP2A	DNA topoisomerase 2-alpha	LUAD	E712V	4FM9A4	OTH
			R736L	4FM9A7	OTH
TTK	Dual specificity protein kinase TTK	LUAD	C604F	2ZMDA1	ENZ
		BRCA	G666E	2ZMDA1	ENZ
XDH	Xanthine dehydrogenase/oxidase	LUAD	C43F	2E1QD3	OTH
			N461T	2E1QD8	OTH

In Binding Site

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	3	0	0	17	0	0	0	0	0	0	0	6	0	0	0	9	5	0	0
C	0	0	0	4	27	0	0	0	0	0	0	0	0	0	29	9	0	0	10	9
D	5	0	15	0	27	3	0	0	0	4	0	0	0	0	0	0	0	1	0	0
E	2	0	11	0	19	0	0	6	0	0	0	0	8	0	0	0	9	0	0	0
F	0	10	0	0	1	0	2	0	14	0	0	0	0	0	11	0	2	0	3	0
G	5	1	7	4	0	0	0	0	0	0	0	0	0	0	11	1	0	2	1	0
H	0	0	9	0	0	0	0	2	0	2	9	23	25	0	0	0	0	0	4	0
I	0	0	0	0	2	0	0	3	13	5	0	0	2	9	8	8	0	0	0	0
K	0	0	0	113	0	0	0	0	0	11	0	13	7	0	7	0	0	0	0	0
L	0	1	0	0	17	0	1	1	0	3	0	13	20	41	11	0	12	8	0	0
M	0	0	0	0	0	0	4	5	8	0	0	0	4	0	13	7	0	0	0	0
N	0	0	39	0	0	0	6	2	23	0	0	0	0	4	5	0	0	5	0	0
P	6	0	0	0	0	3	0	0	8	0	0	3	7	1	2	0	0	0	0	0
Q	0	0	0	25	0	0	7	0	3	6	0	15	24	0	0	0	0	0	0	0
R	0	2	0	0	0	30	7	0	4	0	2	4	7	11	2	0	3	0	0	0
S	12	5	0	0	2	18	0	0	0	0	19	11	0	17	6	0	1	1	0	0
T	29	0	0	0	0	0	2	4	0	2	4	24	8	7	0	0	0	0	0	0
V	29	0	7	8	2	51	0	9	0	13	3	0	0	0	0	0	0	0	0	0
W	0	2	0	0	0	29	0	0	0	0	0	0	0	17	1	0	0	0	0	0
Y	0	7	16	0	1	0	7	0	0	0	0	0	0	0	4	0	0	0	0	0

Surface

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	12	34	0	55	0	0	0	0	0	0	48	0	0	12	56	27	0	0	0
C	0	0	0	12	96	0	0	0	0	0	0	0	0	211	102	0	0	54	104	0
D	41	0	157	0	59	21	0	0	0	12	0	0	0	0	0	0	9	0	10	0
E	28	0	81	0	67	0	0	47	0	0	0	96	0	0	0	17	0	0	0	0
F	1	30	1	0	2	0	24	0	107	0	0	0	0	108	0	41	2	33	0	0
G	24	4	66	63	0	0	0	0	0	0	0	0	0	86	22	2	21	7	0	0
H	0	0	158	0	0	0	0	11	0	10	90	159	223	0	0	0	0	33	0	0
I	0	0	0	0	17	1	0	13	44	133	29	0	0	18	47	50	92	0	0	0
K	1	0	0	559	0	0	0	2	1	9	88	1	116	71	0	30	0	0	0	0
L	0	0	0	1	89	1	43	18	0	20	142	51	377	98	0	107	66	0	0	0
M	0	0	0	0	0	0	82	38	49	0	0	0	48	0	78	74	0	0	0	0
N	0	0	307	0	0	0	69	18	229	0	1	3	0	1	37	32	0	0	17	0
P	40	0	0	0	0	11	0	0	66	0	0	20	82	28	20	0	0	0	0	0
Q	0	0	0	297	0	0	46	0	21	30	0	146	210	0	0	0	0	0	0	0
R	0	22	0	2	0	109	162	1	73	19	9	52	52	47	11	0	24	0	0	0
S	122	21	0	0	14	54	0	6	0	9	0	83	132	0	175	62	0	9	15	0
T	126	0	0	0	0	0	35	34	0	16	11	94	0	73	34	0	0	0	0	0
V	99	0	31	53	13	149	0	62	0	100	41	0	0	0	0	0	0	0	0	0
W	0	5	0	0	0	127	0	0	4	0	0	0	0	158	7	0	0	0	0	0
Y	0	22	167	0	12	0	74	0	0	0	25	0	0	0	37	0	0	1	0	0

Buried

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	1	2	0	27	0	0	0	0	0	0	16	0	0	4	26	33	0	0	0
C	0	0	0	12	45	0	0	0	0	0	0	0	0	20	31	0	0	36	27	0
D	31	0	21	0	34	7	0	0	0	3	0	0	0	0	0	0	12	0	2	0
E	21	0	14	0	29	0	0	20	0	0	0	8	0	0	0	0	12	0	0	0
F	1	32	1	0	0	0	24	0	89	0	0	0	0	0	26	0	46	0	13	0
G	26	4	8	4	0	0	0	0	0	0	0	10	6	0	15	6	0	0	0	0
H	0	0	29	0	0	0	0	19	0	2	19	17	22	0	0	0	0	11	0	0
I	0	0	0	11	0	0	1	37	95	4	0	0	6	17	18	78	0	0	0	0
K	0	0	0	40	0	1	0	3	1	8	18	0	14	7	0	14	0	0	0	0
L	0	0	0	95	2	5	23	0	11	0	40	8	47	32	0	111	33	0	0	0
M	0	0	0	0	0	0	71	2	45	0	0	0	5	0	25	55	0	0	0	0
N	0	0	56	0	0	0	7	14	13	0	0	0	0	19	14	0	0	1	0	0
P	21	0	0	0	0	3	0	0	42	0	0	1	5	2	6	0	0	0	0	0
Q	0	0	0	31	0	0	8	0	1	18	0	24	18	0	0	0	0	0	0	0
R	0	9	0	0	0	49	16	2	4	19	7	0	8	5	25	2	0	9	0	0
S	81	29	0	0	14	33	0	10	0	8	0	19	33	0	20	13	0	7	4	0
T	103	0	0	0	0	0	39	4	0	13	8	15	0	2	15	0	0	0	0	0
V	95	0	5	4	9	56	0	57	0	81	19	0	0	0	0	0	0	0	0	0
W	0	4	0	0	0	57	0	0	0	3	0	0	0	15	1	0	0	0	0	0
Y	0	35	35	0	14	0	28	0	0	0	2	0	0	0	15	0	0	1	0	0

Figure 5.10. Occurrence of individual missense mutations. The counts of missense mutations at the amino acid level divided classified as being adjacent to the binding site, elsewhere on the surface of the protein, or buried in the protein interior. The original amino acid is listed row-wise and the subsequent mutation is listed column-wise.

For small molecules to engage their targets with high affinity, a well-defined cavity that possesses suitable shape and physicochemical properties. The lack of such cavities is partly responsible for the difficulty in developing small-molecule therapeutic agents that bind directly to highly promising cancer targets such as mutated RAS GTPase or transcription factors such as c-MYC. Conversely, the success of kinases as oncology targets can be attributed to the well-defined ATP-binding site. Using binding sites of kinases and other druggable targets, several algorithms have been developed to predict the druggable nature of a binding site using the three-dimensional structure of the protein that harbors them [256]. Among them, SiteScore and DrugScore, which have been developed using data from binding sites occupied by approved drugs [221, 257]. Druggable sites, the highly conserved nature of the ATP-binding site has been the main impediment in the development of kinase drugs. Developing highly selective kinase inhibitors is notoriously difficult, although some successes have been reported. Identifying novel targets with unique druggable binding sites located on potential cancer targets may lead to cancer therapeutics with greater efficacy and lower toxicity.

Here, in an effort to facilitate the chemical probing of new targets in cancer, we explore RNA-seq data of 10 tumor types at TCGA to identify unique and druggable binding sites on proteins encoded by protein products of overexpressed genes. The large-scale effort of TCGA to sequence the genome of tumors from more than 30 cancers provides an unprecedented opportunity to uncover new targets for the development of cancer therapeutics. We identified genes whose mRNA levels are overexpressed in tumors compared with normal tissue. Patient data provided by TCGA was used to further narrow the list of targets to genes whose overexpression correlates strongly with patient survival. This was accomplished by constructing survival curves and evaluating a hazard ratio for each overexpressed gene. Genes with hazard ratio of 1 or greater were considered to correlate with worse patient survival. For each of the 10 diseases that we have considered in this work, we identified protein products of genes whose mRNA levels are differentially expressed that strongly correlate with patient survival. Additionally, we explored these targets in the context of cancer related signaling pathways and the protein-protein interaction network.

The exponentially growing list of three-dimensional structures of proteins prompted us to search the PDB to identify structures for protein products of up-regulated genes that we identified. We used a stringent threshold for these scores to ensure that small molecules that bind to the druggable binding sites have the potential to be developed into therapeutic agents. Among all up-regulated genes we found that 23% of their protein products had a structure at the PDB. Among the 1218 proteins with structures, 405 (33%) had druggable binding sites. A similar ratio was found

among individual diseases. For example, 51 proteins with a crystal structure from among 211 in TNBC had a druggable binding site, while 114 proteins with a crystal structure in LUAD were found to have a binding site among a total of 363. When overexpressed genes are further filtered by hazard ratio, a total of 54 proteins that possess druggable binding sites and 65 possessed binding sites are identified among 1344 differentially expressed genes. There were 15 druggable proteins that are present in multiple tumor types. The most frequently occurring were *MELK* in 4 tumors.

The presence of a binding site is not sufficient to serve as a suitable target site for chemical probe development and drug discovery. The binding site must possess functional relevance. Its position must be located at a site such that the binding of a small molecule will impair the function of the protein harboring the binding site. For example, small molecules that bind to a binding site located at an enzyme active site or protein-protein interface will disrupt enzyme activity or protein-protein interactions and thereby impair the function of the target protein. Binding sites located outside an enzyme active site or protein-protein interface, may or may not modulate the activity of a protein. We classified all binding sites into enzyme active sites, protein-protein interaction sites, or other sites with yet unknown function that may provide an opportunity to modulate protein function through an allosteric mechanism.

Many of the enzyme active sites occur on well-established oncology targets or have been inhibited by small molecules. However, there were several examples of enzymes whose function was explored in cancer but were never targeted with small molecules; these include *PKMYT1*, *STEAP3*, and *NNMT*. There were also several druggable active site binding sites that occurred on enzymes that have seldom been considered in cancer, such as *PYCRI*, *HSPA6*, and *QPRT*. We identified several proteins whose overexpression correlate with patient outcome that occurred at protein-protein interfaces. This discovery is highly significant as protein-protein interactions have been historically challenging due to the lack of well-defined binding sites at protein-protein interfaces [23, 258]. Protein-protein interfaces can offer an opportunity to develop highly selective compounds since many of these interfaces are structurally unique. Among all differentially expressed proteins with binding sites, 18% have binding sites that occurred at protein-protein interfaces. For the proteins encoded by genes that correlate with patient survival, we identified 28 binding sites (7 druggable) on 25 proteins that occurred at protein-protein interfaces. Among these proteins, 13 have been studied in cancer. Examples include *MEF2B*, *HNF4A*, and *CBX2*. The remaining 15 proteins have seldom been studied in cancer, such as *CASC5* and *ZBTB32*. Interestingly, several protein structures possess both PPI and ENZ binding sites either on the same domain (e.g. *GAD1*, *NAMPT*, and *NR1I2*) or on different domains (e.g. *EPHB2*, *PLK1*, and

NTRK1). Small molecules that bind to a binding site on these proteins may serve as allosteric modulator of PPI interactions.

We found that the majority of binding sites were not located either at an enzyme active site or protein-protein interaction site. We refer to these binding sites as other (OTH). Of the 601 unique binding sites on the 202 proteins encoded by genes whose overexpression correlates with patient survival, 102 are ENZ, 46 are PPI, 444 are OTH, and 9 have been classified as both ENZ and PPI. It is likely that many of these OTH binding sites occur at protein-protein interfaces. To explore this possibility, we searched protein-protein interaction databases such as PrePPI for binding partners. Among 759 OTH binding sites located on overexpressed proteins, we identified 17 candidates that have the potential to be located at PPI interfaces. Examples of these proteins include *ANK1*, *CHN1*, and *NCSI*. While OTH binding sites that occur at enzyme active sites or protein-protein interaction sites can be used to develop probes that directly modulates the function of the target harboring these binding sites, the remaining OTH binding sites can provide an opportunity to modulate receptors through an allosteric mechanism [223, 259]. Whether a small molecule that binds to a binding site will allosterically modulate enzyme function or a PPI interaction is difficult to predict. Small molecules can serve as positive or negative allosteric regulators [167, 260, 261]. These OTH binding sites can also be used for the development of small molecules that can be attached to probes for proteasome degradation [262].

Finally, we mapped mutations that were previously identified at TCGA [247] onto the three-dimensional structure of proteins that are encoded by overexpressed genes that correlate with patient outcome. A recent study explored the role of mutations on tumorigenesis [263] and more recently using a structural genomics based approach [264, 265]. Our work complements these studies by identifying druggable binding pockets and classifying pockets into whether they occur at enzyme active sites or protein-protein interaction sites. Mutations that occur within these pockets are expected to have direct consequences to the function of a protein. These pockets could provide promising targets for the development of small-molecule therapeutic agents. Interestingly, several mutations occurred in enzyme active sites. These mutations may either enhance or inhibit enzyme activity. Most of the enzyme mutations appear to involve dramatic changes in physico-chemical properties such as H113Q, G568W, R140L, M80R for *CA6*, *KIFC1*, *NEK2*, and *SULT4A1*. Others involved subtler mutations such as V46A, A287S, and M52T for *CHEK1*, *PCK1*, and *PSPH*, respectively. Since we have focused on proteins that are expected to be overexpressed, it is likely that these mutations will further enhance the active of these enzymes. Three mutations were identified to occur at protein-protein interfaces, R293P, W167L, and Q107H. The first two may have disruptive effects considering that proline residues tend to disrupt secondary structures and

tryptophan residues are generally believed to tighten protein-protein interactions. The overwhelming majority occurred at OTH binding sites. These mutations provide an opportunity to validate the importance of these pockets. It suggests that these pockets may be located at unknown active sites or protein-protein interfaces. Considering that many of these OTH pockets occur on enzymes, it is more likely that they may be located at a protein-protein interface and could be useful targets for the disruption of protein-protein interactions.

5.4 MATERIALS AND METHODS

5.4.1 Gene Expression. Level 3 gene expression data expressed using RNA-seq (RNASeq Version 2) technology for ten cancer types was retrieved from The Cancer Genome Atlas (TCGA). Triple-negative breast cancer (TNBC) patients were identified from a subset of patients in BRCA by filtering clinical records for breast cancer patients who were negative for estrogen receptor (*ER*), progesterone receptor (*PR*), and Her2/neu. The gene expression data was used to build a matrix of read counts for each sample against each mapped gene. Only samples with designations of either the primary solid tumor or the solid tissue normal were kept in this matrix. Differential expression analyses between cancer and normal samples in the RNA-seq expression profiles were conducted using default parameters in the *edgeR* [266] package in R [267]. Differentially expressed (overexpressed) genes were defined as those genes with $p < 0.001$ and $Q < 0.05$. Two \log_2 fold changes of ≥ 2.0 and ≥ 1.5 were used to filter genes for further analysis. Gene symbols provided by TCGA were mapped to their respective UniProt IDs using UniProt's mapping tool (<http://www.uniprot.org/mapping/>).

5.4.2 Protein Structures. An annotated set of 20,192 reference human protein identifiers was retrieved from UniProtKB/SwissProt [227]. The FASTA sequences were retrieved for each of these proteins and used to identify structures in the RCSB Protein Data Bank (PDB) [268]. Each FASTA sequence was queried against the pdbaa dataset using BLASTP (Protein-Protein BLAST v2.2.25+) [269]. To limit the search to protein structures that possess significant sequence identity and coverage to the query sequence, only structures with E-value $< 10^{-5}$, $>90\%$ sequence identity, and PDB sequence coverage $>80\%$ were kept. We then identified the experimental methodology, taxonomy of the identified protein chain, and the structural resolution if the structure was from x-ray diffraction. Previously identified structures were then filtered for only crystal structures from human proteins with a resolution better than 3 Å. To reduce the number of redundant structures identified by BLASTP and generate a representative set of crystal structures associated with each protein, CD-HIT (v4.6.1) [270] was used with default parameters to cluster the FASTA sequences of the PDB structures identified for each of the proteins. Only cluster centers identified by CD-HIT

were used to locate binding sites on the structures for the protein. In total, 4124 proteins had at least one crystal structure that met all of these criteria.

5.4.3 Binding Site Identification. Identification of druggable binding sites on the crystal structures was carried out using the Schrödinger Software Suite. For each cluster identified by CD-HIT, the cluster centers (i.e. the representative structures) were used to identify binding sites. Structures were first retrieved from PDB and binding partners were removed to identify the monomeric representative structures. All other heteroatoms, including solvent molecules and bound ligands, were removed. Selenomethionine residues were converted to methionines. These preprocessed PDB monomeric structures were then processed using the Protein Preparation Wizard workflow. Missing side chains and loops were added with the Prime [271] module. Disulfide bonds were added, and each crystal structure was protonated using PROPKA at pH 7.0. Binding sites were identified using the SiteMap [257] module in Schrödinger on the processed structure. Up to 10 binding sites were kept, while all other parameters were left default. Only binding sites [221] with SiteScore and DrugScore above 0.8 were kept. The average coordinates of the SiteMap spheres were used to identify the centroid of the binding site. Druggable binding sites were distinguished as those with a DrugScore above 1.0. In total, we identified 5498 binding sites on 2607 proteins.

5.4.4 Binding Site Annotation. PyMOL scripts were generated to create individual sessions for each protein with druggable binding sites. The unprocessed protein structure, including all bound ligands and other non-solvent molecules was overlaid back atop the crystal structure. In addition, all redundant structures from the CD-HIT clustering were added and aligned back to the druggable protein. The location of enzymatic binding residues were retrieved from UniProt [227] and Catalytic Site Atlas [228] and highlighted on the processed protein structures.

Each binding site identified by SiteMap was visually inspected and manually annotated to determine its functional role in the protein. If an enzymatic residue was in contact with the SiteMap spheres, or if an enzymatic molecule or inhibitor occupied the space of the spheres, the binding site was labeled ‘enzymatic’ (ENZ). If the binding site was at a protein-protein interaction (PPI) interface on the original structure or on any of the aligned structures, the binding site was labeled ‘PPI’. Otherwise, if the binding site was neither enzymatic nor part of the interaction interface, it was labeled ‘Other’ (OTH). Binding sites of the recognition site of human leukocyte antigens (HLAs) and heme cofactor binding site of Cytochrome P450s were labeled ‘Other’.

Secondary structures for each of the binding sites and their interaction partners were retrieved from DSSP [272]. The secondary structure of each residue of a crystal structure are classified into helix, sheet, or coil in DSSP. The number of residues falling into each category was

retrieved for the residues within 5 Å of the binding site. If there is at least a 60% consensus in the secondary structures for these residues, it was assigned into that category. Otherwise, the binding site was considered mixed.

5.4.5 Survival Analysis. Kaplan-Meier curves were built using the *survival* package in R. For each disease, each patient's time to last follow-up or time to death was collected from the clinical data depending on whether or not the patient was deceased. A patient's overall survival was paired with their respective \log_2 CPM and for diseases using RNA-seq. Expression levels for each gene was separated into 'high expression' and 'low expression' groups using the median expression of the gene across all patients for a given disease. A Cox proportional hazards regression model was fitted to the survival profile to determine the hazard ratio (HR) of each gene. Genes were filtered using $p < 0.05$ and $HR > 1.0$.

5.4.6 Signaling Pathway. 27 cancer related signaling pathways were collected from KEGG [239]. Individual proteins within each of these pathways were collected and mapped to their respective UniProt IDs using the REST API in KEGG. Any protein that could not be mapped to a UniProt entry from the reference protein identifiers was filtered out.

5.4.7 Protein-Protein Interaction Network. A protein-protein interaction network was constructed using the NetworkX [273] module in Python by retrieving human PPI data with experimental evidence from seven major interaction databases: Biomolecular Interaction Network Database (BIND) [274], BioGRID [275], Database of Interacting Proteins (DIP) [276], Human Protein Reference Database (HPRD) [277], IntAct [278], Molecular INTeraction database (MINT) [279], and Reactome [280]. Only those interactions with at least two occurrences among the seven databases were kept. The resulting network featured 9665 nodes and 38164 edges.

5.4.8 Missense Mutations. Mutations were obtained from a recent study by Kandath and coworkers [247]. The work identified somatic variants from 12 cancers as part of TCGA's Pan-Cancer initiative. We only use missense mutation data as other mutations result in the insertion or deletion of amino acids from the protein sequence, which would be very difficult to model onto the three-dimensional structure of the protein. Mutations were mapped using the sample ID barcode provided by TCGA to match patients with both mutation and gene expression data. The data for three diseases were not used since THCA was not included in the original study, while COAD and UCEC had low numbers of patient samples with matched gene expression data. Genes were mapped from Ensembl Transcript IDs to UniProt IDs using UniProt's mapping tool. For each protein, the subsequent amino acid position on the protein sequence was mapped to the protein structure using the pairwise function in BLASTP. Each mutation was then classified by minimizing the Euclidean distance from the corresponding alpha carbon of the mutated residue to the site points

(grid spheres) of each binding site on the protein structure. In addition, the solvent-accessible surface area (SASA) of the mutated residue was calculated using NACCESS [281]. We used the SASA and distance to the closest binding site to classify each mutation as being (i) adjacent to a binding site; (ii) elsewhere on the protein surface; or (iii) buried in the interior of the protein. If the distance between the mutation and the closest binding site was less than 4 Å, the mutation was classified as being adjacent to the binding pocket. Otherwise, if the SASA of the mutated residue was greater than 10 Å², the mutation was classified as being on the surface of the protein. If the mutation did not fit into either of these criteria, it was classified as located in the interior of the protein.

Chapter 6

TUMOR-SPECIFIC CHEMOGENOMIC LIBRARIES BY STRUCTURE-BASED ENRICHMENT FOR GLIOBLASTOMA PHENOTYPIC SCREENING

6.1 INTRODUCTION

Like most solid incurable tumors, glioblastoma multiforme (GBM) exhibit multiple hallmarks of cancer as delineated by Hanahan and Weinberg [1]: Self-sufficiency in growth signals, insensitivity to growth inhibitory signals, evasion from programmed cell death (apoptosis), ability to undergo limitless cycles of cell growth, sustained ability to be supplied by blood (angiogenesis), and aggressive invasion of the brain parenchyma. These phenotypes are driven by multiple targets spanning interconnected signaling pathways across the human protein-protein interaction network. Large-scale sequencing studies of human tumors such as The Cancer Genome Atlas (TCGA) project have revealed that the complex phenotypes that define cancer are driven by a large number of somatic mutations that occur in proteins across the cellular network [216]. Recent whole genome sequencing studies have profiled the molecular signatures of various cancers, including ovarian [4], colorectal [5], breast [6], renal [7], lung [3, 8, 9], pancreatic [10, 11], and brain [12, 13], to identify underlying driver mutations and gene signatures of each disease. These studies have been instrumental not only in classifying tumors and uncovering genetic alterations in cancer cells (mutations, copy number, and rearrangements), but also a comprehensive resource for identifying potential targets.

There is growing interest in harnessing the vast amount of tumor genomic data to guide phenotypic screening and cancer drug discovery. The discovery of small molecules that selectively suppress multiple targets and signaling pathways will likely have greater efficacy in the treatment of these tumors if molecules can be identified that selectively target tumor cells and not normal cells. There has been a resurgence of interest in phenotypic screening in cancer drug discovery [282]. Between 1999 and 2008, over half of FDA-approved first-in-class small-molecule drugs were discovered through phenotypic screening [283]. The increased interest in phenotypic screening is due in part to the lack of effective treatment options for incurable tumors such as GBM, which remains the most aggressive brain tumor and responds poorly to standard-of-care therapy that includes surgery, irradiation, and temozolomide. Standard-of-care therapies for GBM have been essentially unchanged for decades with a median survival of only 14-16 months and a five-year survival rate of 3–5% [284, 285]. Ineffective tumor cell killing is largely due to intra-tumoral genetic instability which allows these malignancies to modulate cell survival pathways, angiogenesis, and invasion [286, 287]. In addition, the highly immunosuppressive GBM

microenvironment complicates therapeutic approaches that minimize tumor burden and promote host immunity [288-290]. Moreover, investigations to date indicate therapies that combine TMZ with immunotherapy-based approaches can either promote or deplete immunity [291-294]. Phenotypic screening can be an effective strategy for the development of small molecules to perturb the function of proteins that drive tumor growth and metastasis. Despite the increased interest in phenotypic screening in cancer drug discovery, the main limitations of the approach include (i) the lack of methods to tailor library selection to the tumor genome, (ii) cellular assays that do not accurately represent a tumor, (iii) overreliance on immortalized cell lines, (iv) targeting a single protein when tumors are driven by multiple proteins, and (v) confining compound screening to one phenotype.

To date, most phenotypic screens are carried out on well-annotated tool compound libraries that include FDA-approved drugs. These are known as chemogenomic libraries, and they are used to uncover new biology for targets associated with these compounds or for drug-repurposing purposes [295-298]. However, existing approved drugs and tool compounds act on less than 5% of targets in the human genome [39]. The lack of target diversity in chemogenomic libraries presents an opportunity for the development of new chemical libraries or for the enrichment of existing libraries. The creation of diverse libraries for high-throughput screening is a major challenge considering the vastness of chemical space. Just among commercially-available compounds, there are now at least 400 million small organic compounds that can be purchased [208, 299]. In addition, specialized libraries designed using diversity-oriented synthesis (DOS) [300] and *de novo* combinatorial libraries such as SCUBIDOO [199] offer additional avenues to screen unexplored chemical space.

Traditional two-dimensional monolayer assays utilizing cancer cell lines have been the most practical method to phenotypically screen these large libraries. However, these screening campaigns have yielded compounds that fail to model compound efficacy and cytotoxicity in more disease-relevant assays [301]. Traditional two-dimensional assays do not accurately capture the three-dimensional microenvironment of tumors, thus leading to toxic compounds that generally tend to block microtubule dynamics or lead to DNA modification [302, 303]. The use of immortalized cell lines to predict efficacy is now also recognized to be inadequate [304]. There are now many examples of small molecules that are efficacious in traditional *in vitro* and *in vivo* models yet fail to show clinical efficacy [305]. As a result, there has been intense interest in the development of more sophisticated three-dimensional assays. Cancer cells grown in three-dimensional spheroids are now widely used to investigate the effects of small molecules on tumor growth and other endpoints such as invasion and remodeling of the tumor matrix. More

sophisticated assays, such as spheroid and organoids, have been developed to better represent the tumor and its microenvironment.

Here we follow a rational approach to create chemogenomic libraries that are used for phenotypic screening to uncover novel GBM targets and generate starting points for the development of GBM therapeutic agents. We follow an innovative approach that combines catalogs of differentially expressed molecular targets identified by tumor genomic profiles along with cellular protein-protein interaction data to select a collection of targets with druggable binding pockets. A total of approximately 9000 in-house compounds are docked to each of these targets. Small molecules that are predicted to simultaneously bind to multiple proteins are selected for phenotypic screening using three-dimensional spheroids of patient-derived GBM cells. Hit compounds that inhibit cancer cell growth are also tested in non-transformed primary normal cell lines in (i) three-dimensional assays using CD34⁺ progenitor cells and (ii) two-dimensional assays using astrocytes. The effect of hit compounds on angiogenesis are also tested using a tube formation assay with brain endothelial cells. To uncover potential mechanisms of action, two compounds were selected for RNA sequencing of compound-treated and untreated cells. Thermal proteome profiling was performed for one compound to identify potential targets. Cellular thermal shift assays using antibodies were used to confirm the binding of the compound to targets that emerged from the thermal proteome profiling study.

6.2 RESULTS

6.2.1 Target Selection, Virtual Screening, and Rank-Ordering of Chemical Library.

A weakness of current implementations of phenotypic screening is the lack of rational approaches in the creation of chemical libraries. Here, we propose a strategy that uses the tumor's genomic profile to enrich chemical libraries for phenotypic screening (**Fig. 6.1**). The process begins with the identification of druggable pockets on a large number of protein structures obtained from the Protein Data Bank (PDB) [306]. In previous work, we searched for druggable binding sites on proteins implicated in a range of cancers and classified them in the context of functional importance [307]. Druggable binding sites were classified based on whether they occurred at a catalytic site (ENZ), a protein-protein interaction interface (PPI), or an allosteric site (OTH).

Using our approach, druggable binding sites were identified for GBM. Gene expression profiles were collected for GBM patients from TCGA. A total of 169 GBM tumors and 5 normal samples have been characterized using RNA sequencing platforms. The data were used to perform differential expression analysis to identify genes that are overexpressed in GBM ($p < 0.001$, FDR < 0.01 , and \log_2 fold change (\log_2FC) > 1) (**Fig. 6.2**). In addition, a set of somatic mutations was

retrieved from GBM patients at TCGA and identified for 158 of the 169 tumor samples. In total, 755 genes with somatic mutations were overexpressed in GBM patient samples.

The set of 755 genes were subsequently filtered based on whether their protein products are involved in protein-protein interactions. Two large-scale protein-protein interaction networks of the human proteome were recently described by Rolland and co-workers [308]. The first network is from literature curation of seven widely used protein-protein interaction databases. The second network is based on systematically mapping human binary protein-protein interactions. The two protein-protein interaction datasets from both the literature-curated and experimentally-determined networks were combined to form a large-scale protein-protein interaction network consisting of approximately 8000 proteins and 27000 interactions. The protein products of the genes implicated in GBM were mapped onto this protein-protein interaction network to construct a GBM subnetwork (**Fig. 6.3A**). Among the 755 previously identified genes implicated in GBM, 390 had at least one interaction in the network. In total, 117 of the 390 proteins had at least one druggable binding site (**Fig. 6.3B**).

To identify small molecules that inhibit phenotypes associated with GBM, an in-house library of approximately 9000 compounds was docked to the set of 316 druggable binding sites on proteins in the GBM subnetwork. The Support Vector Machine-Knowledge Based (SVR-KB) [309] scoring method was used to predict the binding affinities of each pair of protein-compound interactions. The number of druggable binding sites with affinities better than a given SVR-KB cutoff was used to rank-order compounds (**Fig. 6.4A**). In this work, an SVR-KB cutoff corresponding to a computational binding affinity of 10 nM was used. Approximately 55 percent of compounds were predicted to bind to five or less binding sites, and 20 percent of the docked compounds were predicted to bind to none of the pockets in the GBM subnetwork (**Fig. 6.4B**). Less than 4 percent of the compounds were predicted to bind to at least 30 of the 316 binding sites. Two separate phenotypic screens were carried out using different criteria on the number of predicted GBM targets. In the first phenotypic screen, small molecules with the highest number of predicted GBM targets were selected for further testing. In the second phenotypic screen, small molecules were identified using a set cutoff of 10 to 20 GBM targets. A total of 154 compounds were selected for the first screen that were predicted to target between 38 and 86 binding sites on GBM-specific proteins. The top 154 compounds were hierarchical clustered using chemical similarity. For each cluster, the compound corresponding to the cluster center was selected for phenotypic testing.

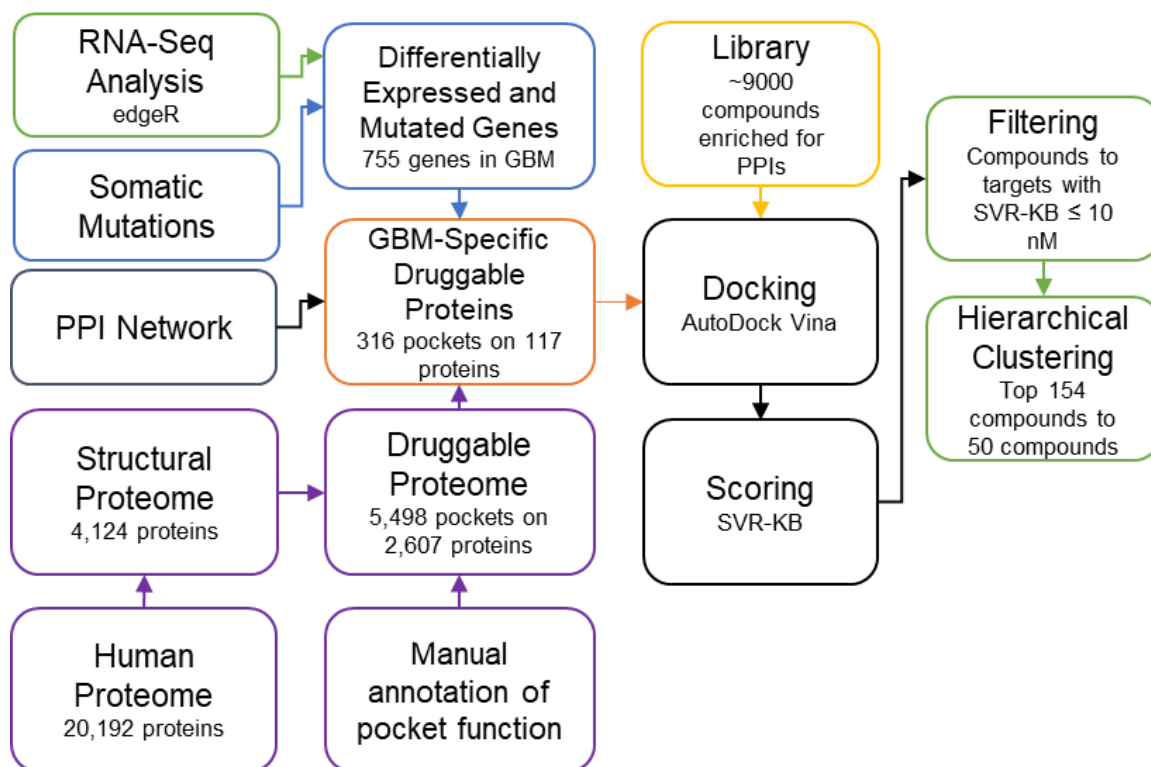


Figure 6.1. Workflow for the identification of druggable targets implicated in GBM. Workflow used to identify GBM-specific druggable targets through integration of genomic RNA-seq, somatic mutation, protein-protein interaction network, and protein structure data. Small-molecule compounds were screened against these pockets to identify compounds that could target proteins implicated in GBM.

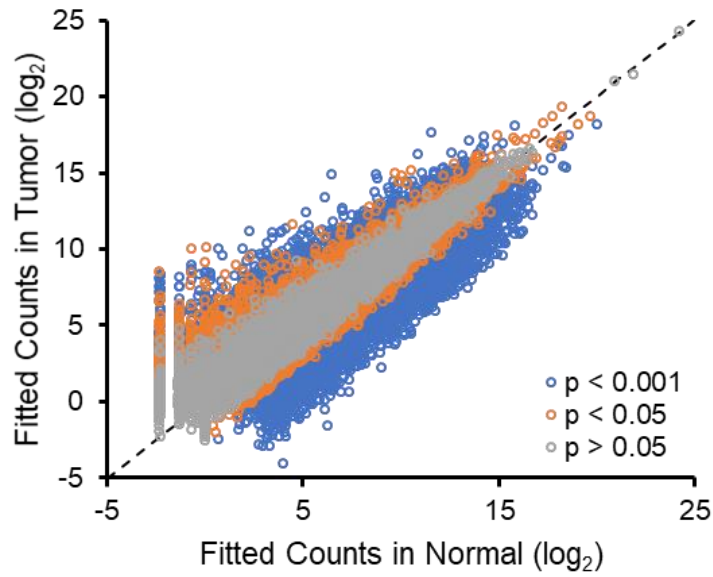


Figure 6.2. Differential expression analysis of 169 tumor and 5 normal GBM RNA-seq samples from TCGA. Mean fitted counts for each gene are shown on the x- and y-axis, respectively.

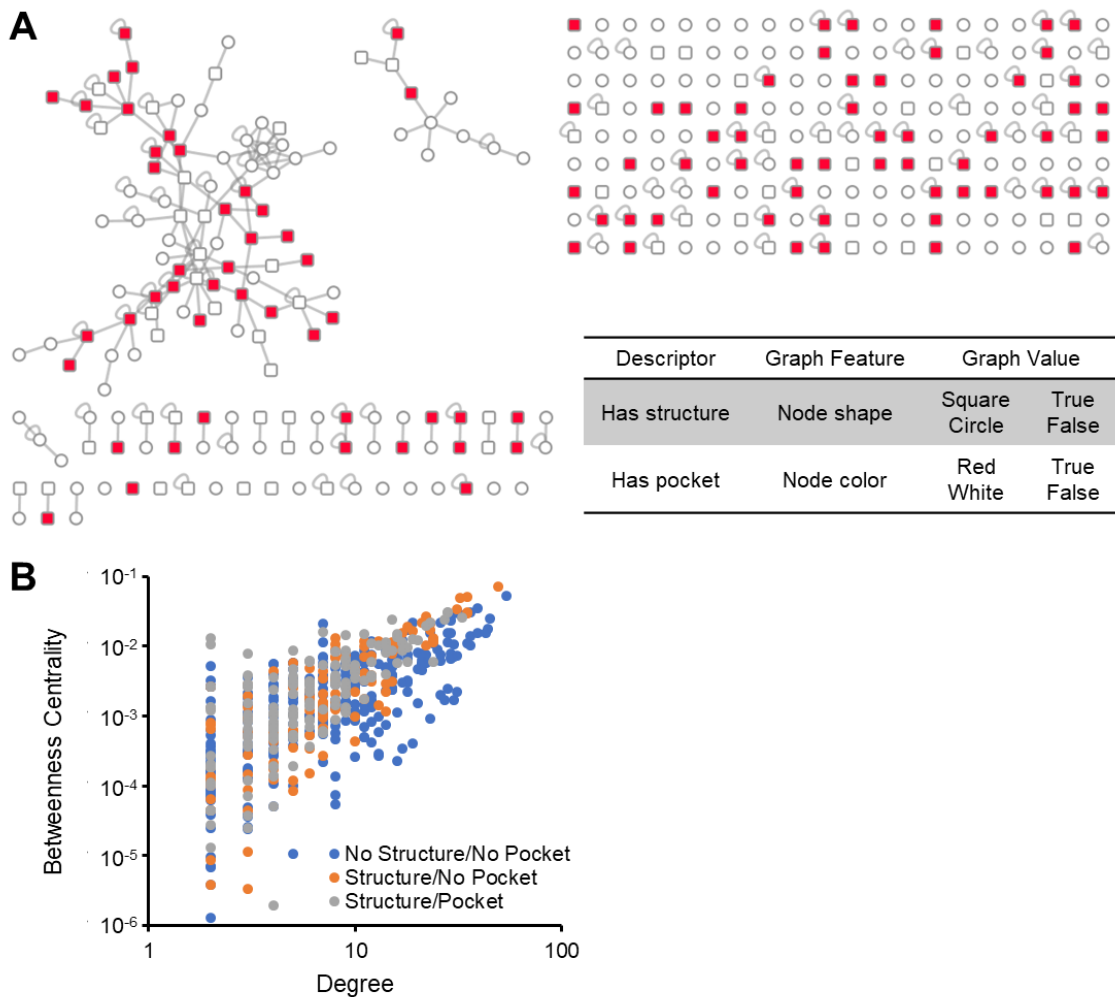


Figure 6.3. Protein-protein interaction subnetwork of GBM-specific targets. **(A)** Interaction network for the GBM-specific targets. Proteins are shown as squares if there is a solved human crystal structure available or as circles otherwise. Proteins with a druggable binding pocket on an associated structure are colored red or white otherwise. **(B)** Degree (number of edges) versus betweenness centrality for GBM-specific targets in the GBM-specific subnetwork.

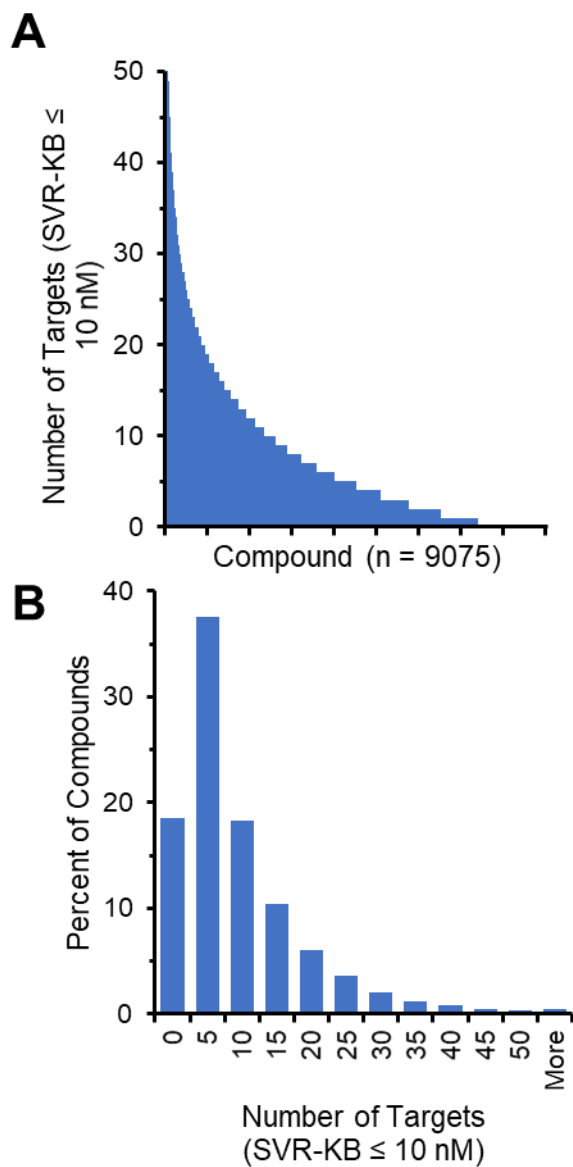


Figure 6.4. Rank-ordering of compounds by their predicted number of predicted GBM-specific targets. **(A)** Waterfall plot of each compound's number of predicted GBM-specific targets using an SVR-KB cutoff of 10 nM. **(B)** Histogram plot showing the percentage of compounds predicted to bind to the number of GBM-specific targets using an SVR-KB cutoff of 10 nM.

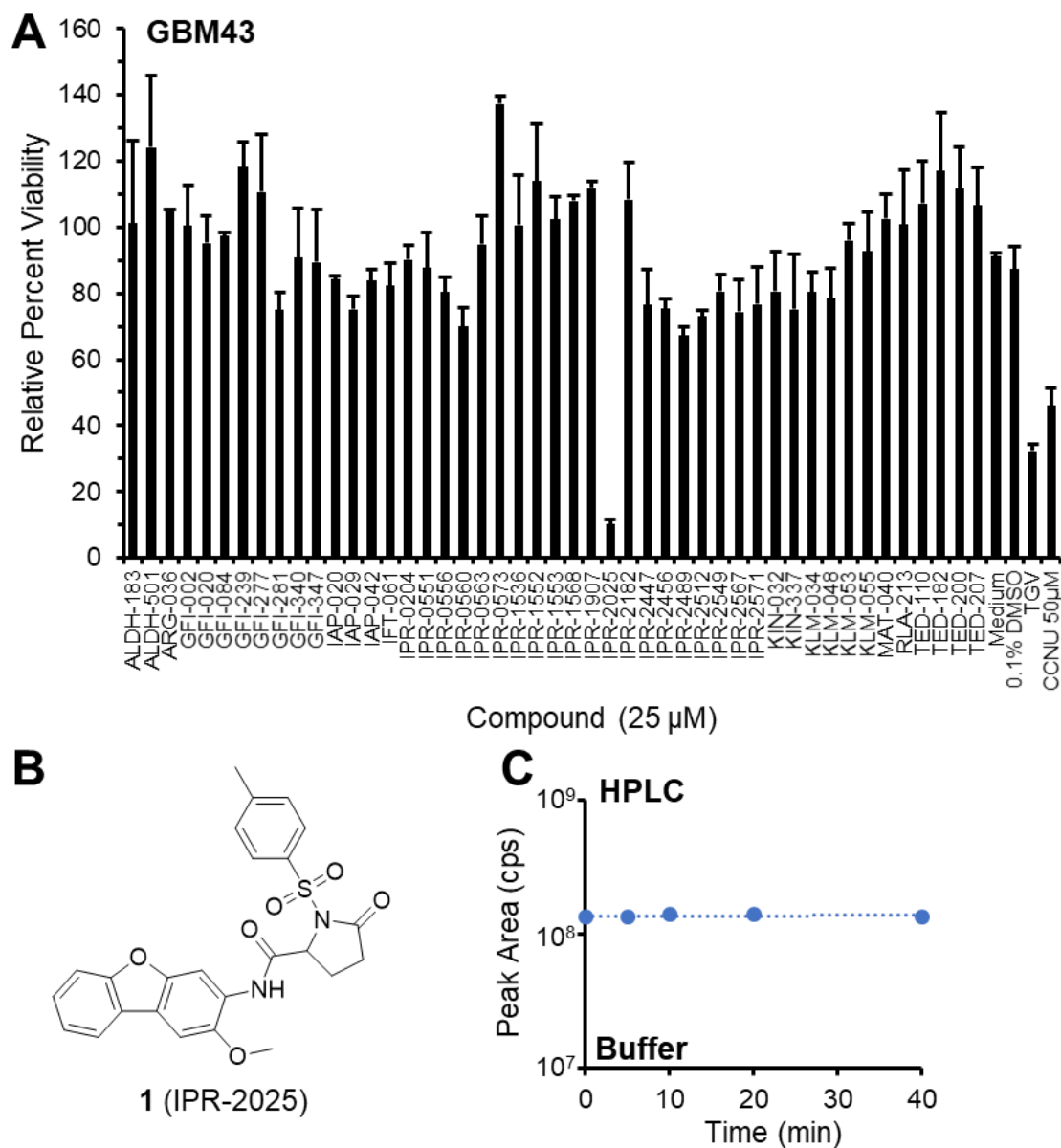


Figure 6.5. Screening compounds against GBM43 leads to the identification of **1** (IPR-2025). **(A)** Screening the compounds predicted to maximally target GBM-specific proteins against GBM43 spheroids at an initial concentration of 25 μ M (mean \pm SD; $n = 3$). Culture medium and 0.1% DMSO were used as negative controls. A TGV cocktail (200 μ M temozolomide, 12.5 μ M ipatasertib/GDC-0068 [pan AKT inhibitor], and 12.5 μ M voxtalisib/VOX [PI3K/mTOR inhibitor]) and CCNU (50 μ M lomustine) were used as positive controls. **(B)** Compound structure of **1** identified from initial screening. **(C)** Stability from incubations of **1** in buffer (0.5 mL of a 1 mL incubation) using high performance liquid chromatography (HPLC). The compound is stable in buffer.

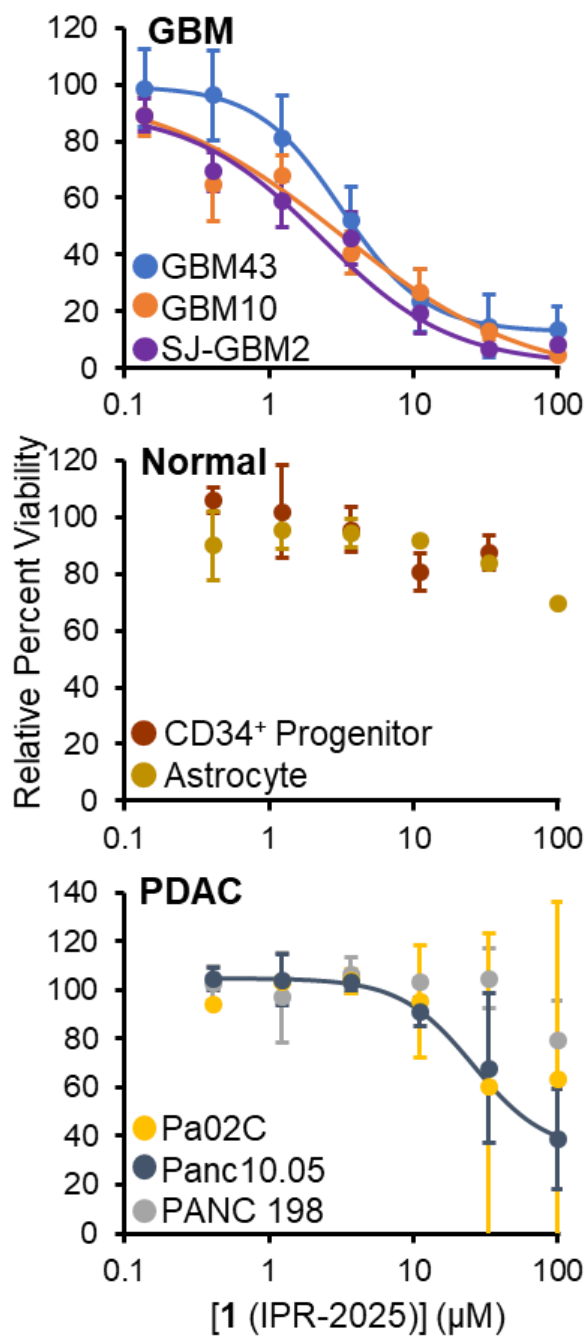


Figure 6.6. Concentration-dependent studies of **1** (IPR-2025) against cancer models. Concentration-dependent screening of **1** against a variety of glioblastoma multiforme (GBM), normal, and pancreatic adenocarcinoma (PDAC) models (mean \pm SD; n = 3).

Table 6.1. Synthesized derivatives of **1** (IPR-2025).

Compound	Structure	Physicochemical Properties				IC ₅₀ (μM) ^a		
		MW	logP	PSA	logBB	GBM43	GBM10	SJ-GBM2
1 (IPR-2025)		478.5	3.6	58.5	-1.00	3.7 ± 0.1	4.9 ± 1.1	2.3 ± 1.2
22 (IPR-3502)		464.5	4.1	58.4	-0.59	NI	NI	ND
23 (IPR-3503)		428.5	4.1	56.5	-0.38	NI	NI	ND
24 (IPR-3504)		324.3	1.8	39.9	-0.85	NI	NI	ND
25 (IPR-3593)		532.5	4.1	58.2	-0.65	28.6 ± 10.4	10.9 ± 4.4	19.6 ± 4.4
26 (IPR-3594)		513.0	4.3	60.4	-0.96	NI	NI	NI
27 (IPR-3595)		513.0	4.3	60.4	-0.96	NI	NI	NI

^aRepresentative of at least two independent experiments, where each concentration point is measured in duplicates (mean ± SD).

ND: Not determined

NI: No inhibition

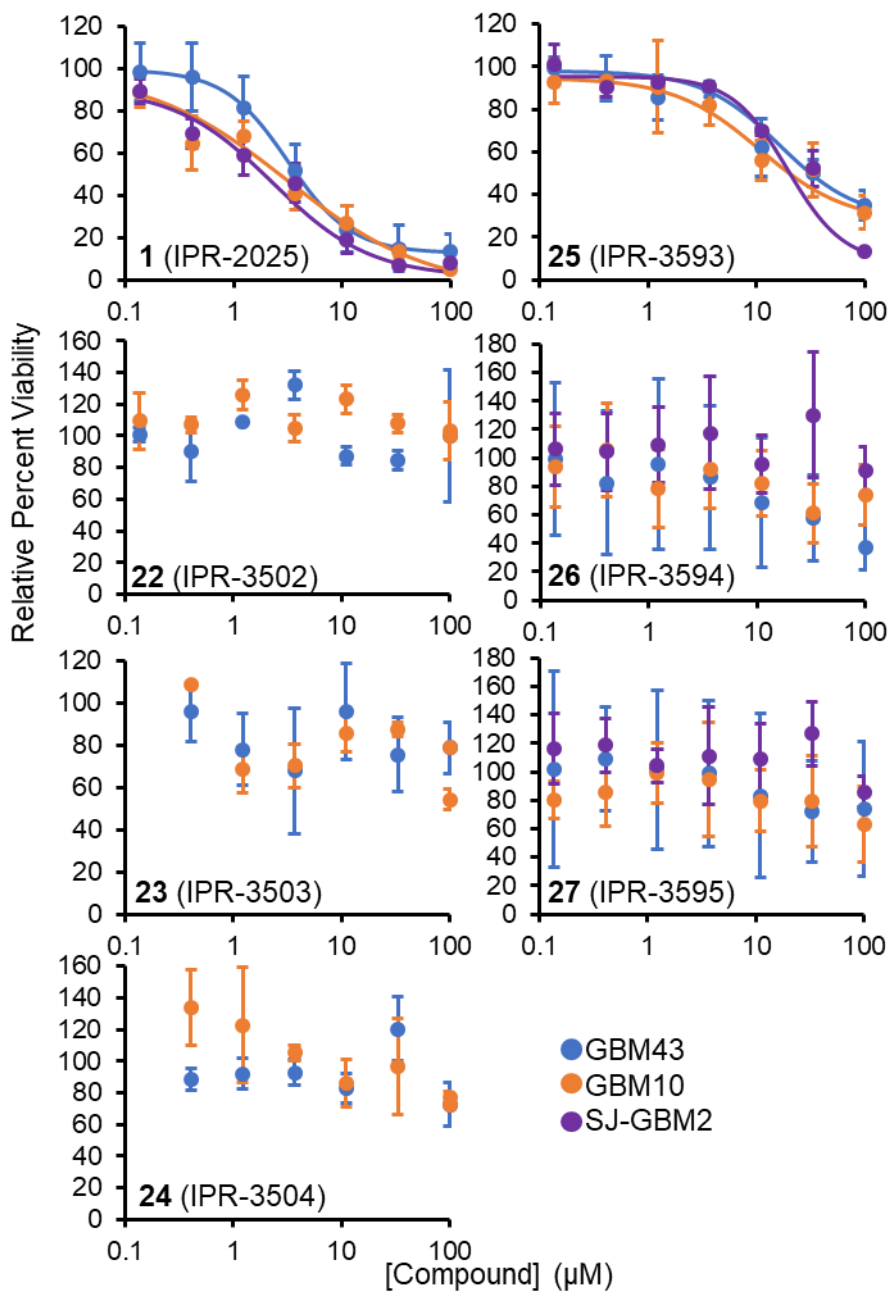


Figure 6.7. Compound activities of resynthesized **1** and synthesized derivatives. Compound activity of **1** and six synthesized derivatives in the three GBM cell lines GBM43, GBM10, and SJ-GBM2.

6.2.2 Exploring Compounds in Patient-Derived GBM Spheroids. A total of 47 compounds selected from the library of 9000 compounds were tested in a three-dimensional spheroid viability assay [310, 311] using the patient-derived primary glioma cell line GBM43 [312] at 25 μM (**Fig. 6.5A**). Most compounds showed little to no effect on GBM43 cell viability, except for **1** (IPR-2025), which inhibited cell viability by 90% (**Fig. 6.5B**). The compound was resynthesized, and its stability was tested following incubation in buffer. The compound was stable in buffer. A follow-up concentration-dependent study of synthesized **1** revealed an IC_{50} of $3.7 \pm 0.1 \mu\text{M}$ in GBM43 (**Fig. 6.6**). Compound **1** was tested in two additional glioblastoma spheroid models derived from different patients with recurrent GBM, namely GBM10 and SJ-GBM2. GBM10 [312] is derived from an adult patient, while SJ-GBM2 [313] is derived from a pediatric patient. The IC_{50} of **1** was $3.1 \pm 1.5 \mu\text{M}$ and $2.9 \pm 1.9 \mu\text{M}$ in GBM10 and SJ-GBM2, respectively. The effect of **1** on normal and non-transformed cell viability was explored using CD34^+ progenitor cells and astrocytes (**Fig. 6.6**). Cell viability studies for CD34^+ cells were performed using a colony formation assay, while a monolayer assay was used for astrocytes. Compound **1** had no effect on the cell viability of CD34^+ or astrocyte cells up to 100 μM . As a clinical comparator in this assay, the standard-of-care temozolomide inhibits GBM43 with an IC_{50} of $244 \pm 24 \mu\text{M}$ [314]. However, GBM43 is known to be moderately resistant to temozolomide [315]. An alternative treatment option is the chemotherapeutic CCNU (lomustine), which inhibited GBM43 and GBM10 cell viability weakly, showing about 50% inhibition at 100 μM .

The activity of **1** was also assessed in three pancreatic ductal adenocarcinoma patient-derived spheroid models [316-318]: Pa02C, Panc10.05 (Pa16C), and Panc198 (Pa20C). Pa02C is from liver metastasis of pancreatic cancer, while Panc10.05 and Panc198 are from the primary pancreatic tumor. Compound **1** showed no activity in Pa02C and Panc198 and weak activity in Panc10.05 with an IC_{50} of $26.0 \pm 5.9 \mu\text{M}$.

6.2.3 Structure-Activity Relationship (SAR) of 1 (IPR-2025). A set of 5 analogs from our internal library were identified with high Tanimoto similarity to **1**: **2** (GFI-027), **3** (IPR-1909), **4** (RAG-021), **19** (IPR-2024), and **20** (KLM-017). These 5 compounds were tested in a concentration-dependent manner in GBM43. Only **19** inhibited in a concentration-dependent manner like **1**, although the compound only reaches 60% inhibition at 100 μM in GBM10 and has an approximate IC_{50} of 33 μM and no effect in GBM43. An analog-by-catalog approach was followed to identify another 15 derivatives. These compounds were tested in GBM43 and GBM10 in a concentration-dependent manner. Substitution of the fused tricyclic moiety in **14** (IPR-3440), **20**, and **21** (IPR-3442), among others, led to loss of activity. Among the derivatives that share the fused tricyclic moiety of the parent, two also feature the sulfonyl group. **5** (IPR-3474) substitutes

the toluene group with a biphenyl and lacks the carbonyl group on the pyrrolidinone group, while **6** (IPR-3476) replaces the pyrrolidinone group with an amine linker. These two compounds suggest the importance of the carbonyl group on the pyrrolidinone. Similarly, compounds that lack the sulfonyl group, like **3**, also lacked activity. To further assess the importance of these moieties, three additional compounds, **22** (IPR-3502), **23** (IPR-3503), and **24** (IPR-3504), were synthesized that lack key moieties (**Table 6.1** and **Fig. 6.7**). The removal of the carbonyl group on the pyrrolidinone, sulfonyl linker, or both the sulfonyl linker and the methylbenzene group in **22**, **23**, and **24**, respectively, led to loss of activity in both GBM43 and GBM10. Three additional derivatives were synthesized to add halogen groups to **1**. Substitution of the methyl group in R₃ with a trifluoromethyl in **25** (IPR-3593) resulted in an almost ten-fold decrease in IC₅₀ across each of the GBM spheroids compared to **1**. Chlorine atoms were added to two separate positions of the fused ring structure in **26** (IPR-3594) and **27** (IPR-3595), resulting in no activity in either compound across the GBM spheroids. Of the 21 analogs of **1** that were tested across each of the three GBM spheroids, only the trifluoromethyl **25** resulted in appreciable IC₅₀s.

6.2.4 Additional Phenotypic Screens with Candidates with Fewer Predicted Targets.

In the previous screen, compounds with the largest number of predicted GBM targets were selected for experimental validation. We repeated this process to explore whether fewer predicted targets would also yield active compounds. Compounds that were selected were predicted to bind to 10-20 GBM targets using a similar SVR-KB cutoff of 10 nM. Among the top 50 compounds, three inhibited GBM43 viability by more than 80% at 25 μ M (**Fig. 6.8A**). The structures of the three hits, **28** (ALDH-22), **29** (IPR-196), and **30** (IPR-1964) are shown in **Fig. 6.8B**.

Each compound was tested in a concentration-dependent manner across the patient-derived GBM spheroid models GBM43 and GBM10 (**Fig. 6.9**). Compound **28** inhibited cell viability of GBM43 and GBM10 with IC₅₀s of 5.7 ± 2.6 and 21.8 ± 8.0 μ M, respectively. Compound **29** inhibited GBM43 with an IC₅₀ of 19.7 ± 3.9 μ M and GBM10 with an IC₅₀ of approximately 30 μ M. Like **29**, **30** inhibited both GBM43 and GBM10 with low micromolar IC₅₀s. Neither **29** nor **30** had an effect on CD34⁺ normal cell growth (**Fig. 6.9**). The activity of **29** was also assessed in the three PDAC spheroid models. Compound **29** inhibited cell viability in all three pancreatic models with approximately 10 μ M IC₅₀.

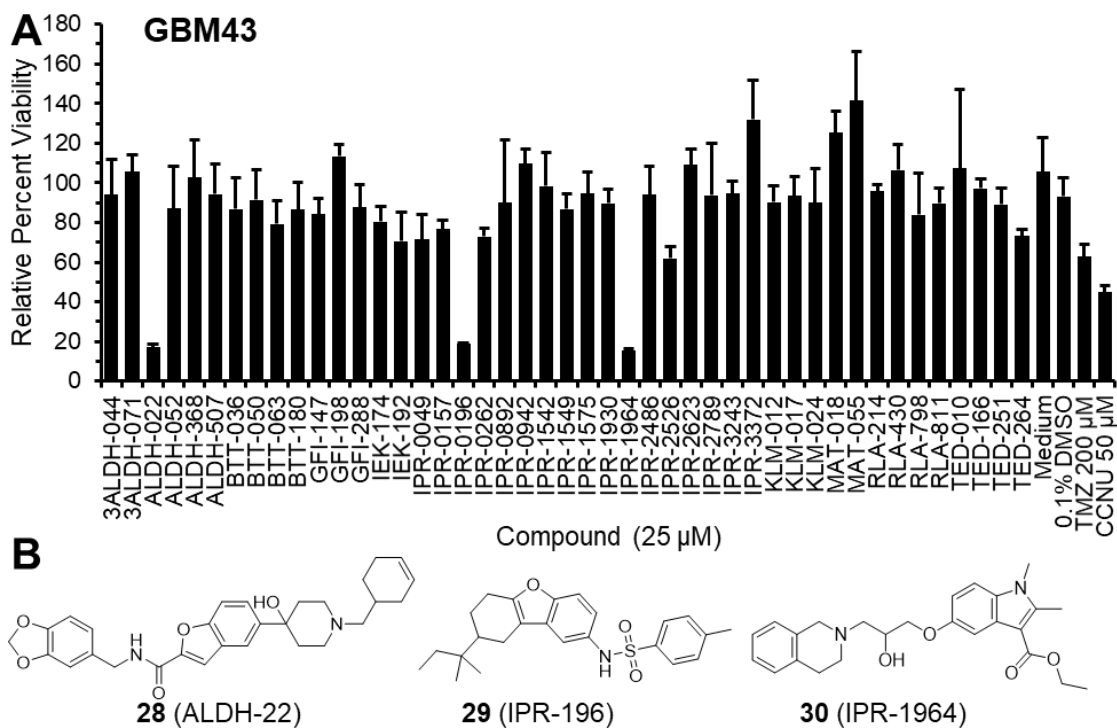


Figure 6.8. Reducing the number of predicted targets to select compounds for exploration in patient-derived GBM spheroids. **(A)** Screening compounds predicted to target between 10 and 20 GBM-specific proteins against GBM43 spheroids at an initial concentration of 25 μ M (mean \pm SD; n = 3). **(B)** Compound structures of three hits identified.

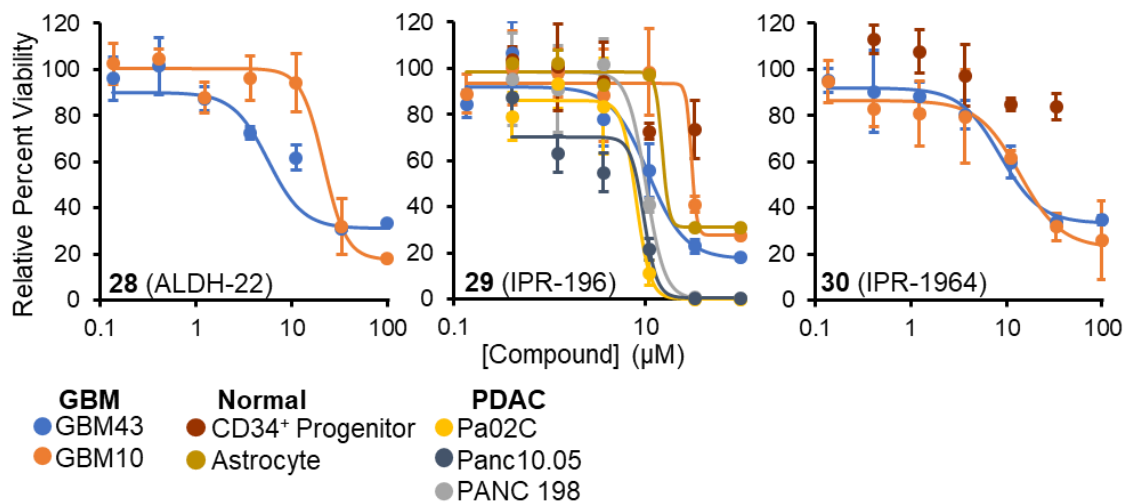


Figure 6.9. Concentration-dependent studies of additional hits against cancer models. Concentration-dependent screening against a variety of GBM, normal, and PDAC models (mean \pm SD; n = 3).

Table 6.2. Predicted protein targets of **1** (IPR-2025).

Symbol	Name	Protein Family	PPI Network		TCGA GBM	
			Degree	Betweenness Centrality	Fold Change	Mutations
<i>Protein-Protein Interaction Interface</i>						
PLK1	Serine/threonine-protein kinase PLK1	Ser/Thr protein kinase	15	1.70E-03	3.3	9
NCF1	Neutrophil cytosol factor 1	-	9	1.70E-04	2.3	6
PNP	Purine nucleoside phosphorylase	PNP/MTAP phosphorylase	2	0.00E+00	1.4	6
<i>DNA-Binding Site</i>						
EXO1	Exonuclease 1	XPG/RAD2 endonuclease	10	3.40E-04	3.7	14
TOP2A	DNA topoisomerase 2-alpha	Type II topoisomerase	3	2.50E-05	7.2	10
<i>Beta-Propeller</i>						
CDC20	Cell division cycle protein 20 homolog	WD repeat CDC20/Fizzy	8	3.10E-04	3.5	9
GNB1	Guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1	WD repeat G protein beta	29	4.60E-03	2.0	9
ITGA5	Integrin alpha-5	Integrin alpha chain	2	1.50E-05	3.0	11
RACK1	Guanine nucleotide-binding protein subunit beta-2-like 1	WD repeat G protein beta	5	1.10E-04	1.4	6
<i>Allosteric Site near Protein-Protein Interaction Interface</i>						
NCF2	Neutrophil cytosol factor 2	NCF2/NOXA1	8	6.50E-04	2.3	9
NEDD4	E3 ubiquitin-protein ligase NEDD4	-	15	1.10E-03	2.0	15
PYGL	Glycogen phosphorylase, liver form	Glycogen phosphorylase	3	0.00E+00	3.2	19
<i>Allosteric Site near Enzymatic Nucleoside</i>						
KIF11	Kinesin-like protein KIF11	Kinesin	4	8.40E-06	2.8	6
TAP1	Antigen peptide transporter 1	ABC transporter	2	2.40E-04	2.4	12

Other Allosteric Site

ACE	Angiotensin-converting enzyme	Peptidase M2	1	0.00E+00	1.9	19
CENPE	Centromere-associated protein E	Kinesin	6	7.50E-05	2.8	17
EZH2	Histone-lysine N-methyltransferase EZH2	Histone-lysine methyltransferase	6	2.00E-04	4.5	10
FLNA	Filamin-A	Filamin	34	7.90E-03	2.5	20
GUSB	Beta-glucuronidase	Glycosyl hydrolase	2	0.00E+00	2.0	7
MMP2	72 kDa type IV collagenase	Peptidase M10A	4	7.40E-04	3.9	12
MMP9	Matrix metalloproteinase-9	Peptidase M10A	4	1.60E-08	7.3	8
NR5A2	Nuclear receptor subfamily 5 group A member 2	Nuclear hormone receptor	6	3.10E-05	4.4	6
NRP1	Neuropilin-1	Neuropilin	2	4.10E-07	1.8	18
PLA2G4A	Cytosolic phospholipase A2	-	1	0.00E+00	1.6	7
SHMT2	Serine hydroxymethyltransferase, mitochondrial	SHMT	2	3.60E-06	2.0	7
SPARC	SPARC	SPARC	1	0.00E+00	3.1	7
TLR2	Toll-like receptor 2	Toll-like receptor	3	4.40E-05	2.0	17

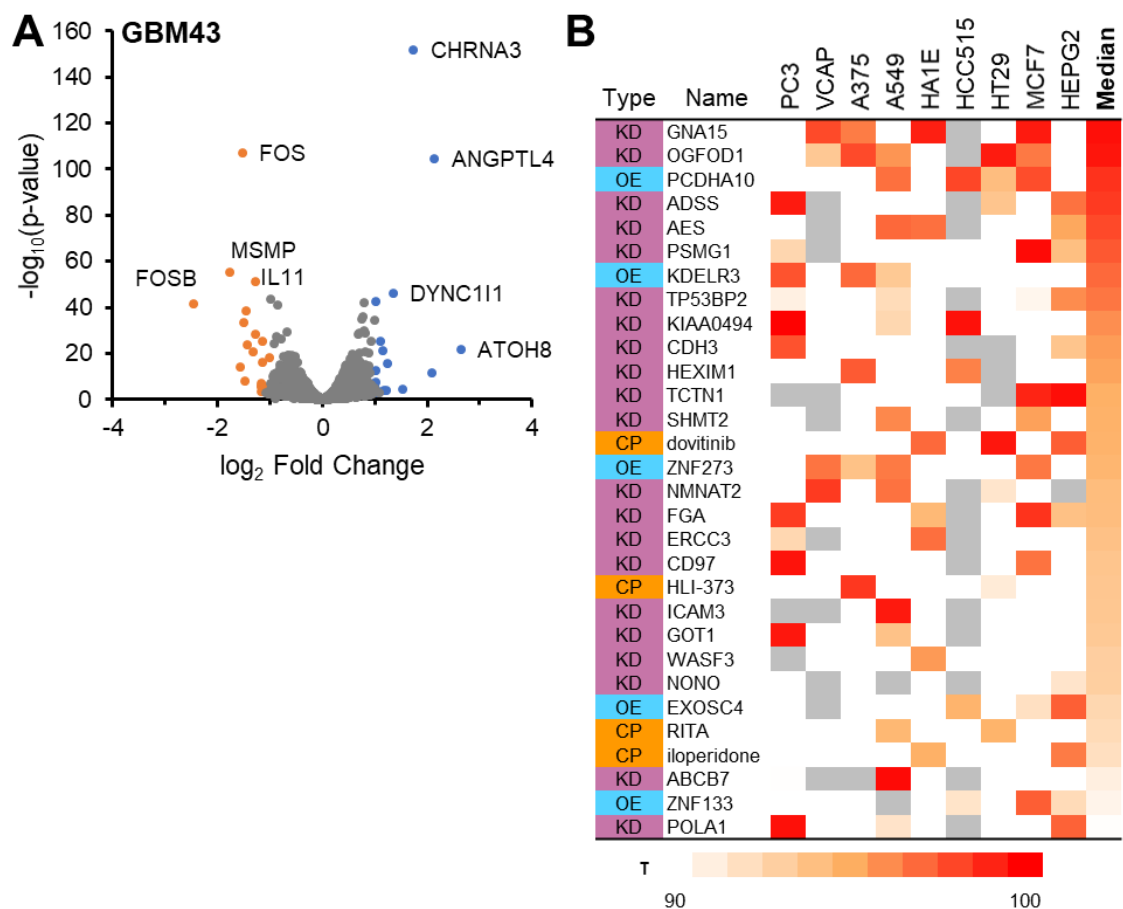


Figure 6.10. RNA-Seq of **1** (IPR-2025) treated GBM43 cells. **(A)** Volcano plot of fold change versus log-transformed significance of GBM43 cells treated with 10 μM of **1** versus control. Differentially expressed genes are identified using a \log_2 fold change cutoff of 1. **(B)** The most similar 30 perturbagens to the gene signature of treated GBM43 identified using the L1000 platform. Perturbagens are rank-ordered using the median τ statistic across nine cell lines, and are classified into gene knockdown (KD), overexpression (OE), and compounds (CP).

6.2.5 Compounds Inhibit Tube-Formation in Matrigel. Since the seminal work of Folkman [319], uncontrolled angiogenesis (the process of new blood vessel growth from existing ones) has become an established hallmark of cancer [1]. Solid tumors require a dedicated blood supply once they reach a certain, limiting size, and antiangiogenic agents can block tumor growth by starving the tumor of oxygen and nutrients. A so-called “angiogenic switch” is an integral part of tumor development: varied tumor types secrete vascular endothelial growth factor (VEGF) and other proangiogenic stimuli and downregulate antiangiogenic proteins. There have also been some promising clinical results with angiogenesis inhibitors in improving progression-free survival in both primary and recurrent GBM [320, 321]. This cancer is also characterized by microvascular proliferation [322] and high levels of VEGF [323]. Given the ample evidence of the importance for angiogenesis in the biology of these tumors, there is significant need to identify novel compounds with specific antiangiogenic activity.

The parent compounds **1**, **29**, **30**, and **31** were tested for their ability to inhibit tube formation of brain microvascular endothelial cells. Compound **1** inhibited tube formation with approximate 0.1 μM IC_{50} , while **29**, **30**, and **31** inhibited tube formation with approximate 1 μM IC_{50} .

6.2.6 Structural Analysis and RNA Sequencing to Uncover Compound 1 Mechanism of Action. The binding modes of **1** to each of the targets predicted by the SVR-KB scoring function were examined in detail. Each target was classified using the structural context of the binding site and functional context of the protein (**Table 6.2**). Generally, the binding modes of **1** were at allosteric sites outside the active site on enzymes. Three allosteric sites were adjacent to known protein-protein interaction interfaces on *NCF2*, *NEDD4*, and *PYGL*. Similarly, **1** was predicted to bind to allosteric sites adjacent to the active sites on *KIF11* and *TAPI*. The compound was predicted to bind at pockets at the protein-protein interfaces at the *PLK1* polo box and at the homomers of *NCF1* and *PNP*. The compound was also predicted to bind to the β -propeller structures of *CDC20*, *GNB1*, *RACK1*, and *ITGA5*.

RNA sequencing was performed on untreated GBM43 cells and GBM43 cells treated with **1** to validate the predicted targets of **1** and uncover a potential mechanism of action (**Fig. 6.10A**). GBM43 cells were treated with **1** and collected for analysis. Differential expression analysis ($p < 0.001$, $\text{FDR} < 0.01$, $|\log_2FC| > 1$) revealed a set of 15 overexpressed and 20 underexpressed genes in GBM43 cells treated with **1**. The sets of overexpressed and underexpressed were separately analyzed for overrepresented GO terms [324]. No significantly overrepresented terms were found among either set of differentially expressed genes. The set of differentially expressed genes (DEGs) were compared with those known to be causally implicated in cancer using the Cancer Gene Census

(CGC) [325]. Two DEGs were previously identified as oncogenes. The transcription factor *GATA2* has been shown to promote GBM progression through the EGFR/ERK/Elk-1 pathway [326]. Similarly, the kinase *KDR* (*VEGFR2*) is a known oncogene in lung, blood, and skin cancers and plays a key role in regulating angiogenesis [327].

The expression profile of cells treated with **1** was compared to the gene signatures of previously characterized compounds and gene knockdowns using the LINCS L1000 platform. The L1000 platform is an extension of the Connectivity Map (CMap) project [243], which used similarities in gene expression signatures to discover shared mechanisms of action between small molecule and genetic perturbations. The L1000 platform expands on CMap and uses a panel of approximately 1,000 landmark genes to characterize the molecular profiles of over 19,000 compound and 5,000 gene perturbations across nine cell lines [328]. To compare the gene expression profile of **1** to the existing signatures in L1000, a signature query was generated using the sets of overexpressed and underexpressed genes. The similarity between the signature query is compared with all other signatures in L1000 using a connectivity score (τ), which corresponds to the percentage of all reference gene sets that are more similar than the observed gene signature. The most similar 30 gene and small molecule perturbations of **1** are shown in **Fig. 6.10B**. The most similar gene signatures are gene knockdowns of *GNAI5* and *OGFOD1*. The G protein α -subunit *GNAI5* is part of the heterotrimeric G protein complex consisting of α -subunit and $\beta\gamma$ complex, which mediate downstream signaling of G protein-coupled receptors (GPCRs) [329]. While no α -subunits were among the targets predicted to bind to the compound, both the G protein β -subunit *GNBI* and β -like subunit *RACK1* were predicted targets of **1**. *RACK1* acts as a scaffolding protein in transcription regulation of the transcription factor *EIF4E* [330]. Gene silencing and knockdown studies of *RACK1* have resulted in promotion of apoptosis and inhibition of cell proliferation, migration, and invasion in glioma [331, 332]. The oxygenase *OGFOD1* belongs to a family of transcriptional factor and chromatin regulators and has been found to inhibit cell proliferation in breast cancer cells [333]. Also among the knockdown gene signatures most similar to the gene signature of **1** in L1000 are the tumor suppressors *AES* [334] and *TP53BP2* [335].

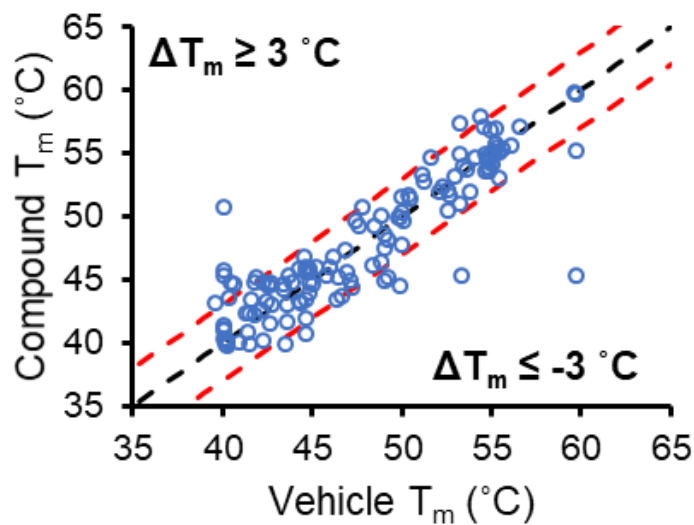


Figure 6.11. Thermal proteome profiling of **1** (IPR-2025) treated GBM43 cells. GBM43 cells treated with 10 μM **1** and untreated cells were serially heated to six different temperatures for thermal proteome profiling. Proteins were identified and quantified using mass spectrometry. Melting curves were fitted for each protein to determine the shift in melting temperature between untreated and treated proteins (ΔT_m). The panel shows a scatterplot of the calculated melting temperature (T_m) in individual proteins between vehicle and compound-treated experiments.

Table 6.3. Proteins with largest increase in melting temperature when treated with compound **1** ($\Delta T_m \geq 3$ °C).

Symbol	Name	ΔT_m (°C)
HTATSF1	HIV Tat-specific factor 1	10.6
SPG20	Spartin	5.6
IK	Protein Red	5.3
DCTN1	Dynactin subunit 1	4.4
CTSB	Cathepsin B	4.1
HDLBP	Vigilin	4.7
TNPO1	Transportin-1	3.7
PPP1CB	Serine/threonine-protein phosphatase PP1-beta catalytic subunit	3.6
EIF3A	Eukaryotic translation initiation factor 3 subunit A	3.3
KIF5B	Kinesin-1 heavy chain	3.3
DNM2	Dynamamin-2	3.3
RACK1	Receptor of activated protein C kinase 1	3.1
PFN2	Profilin-2	3.1

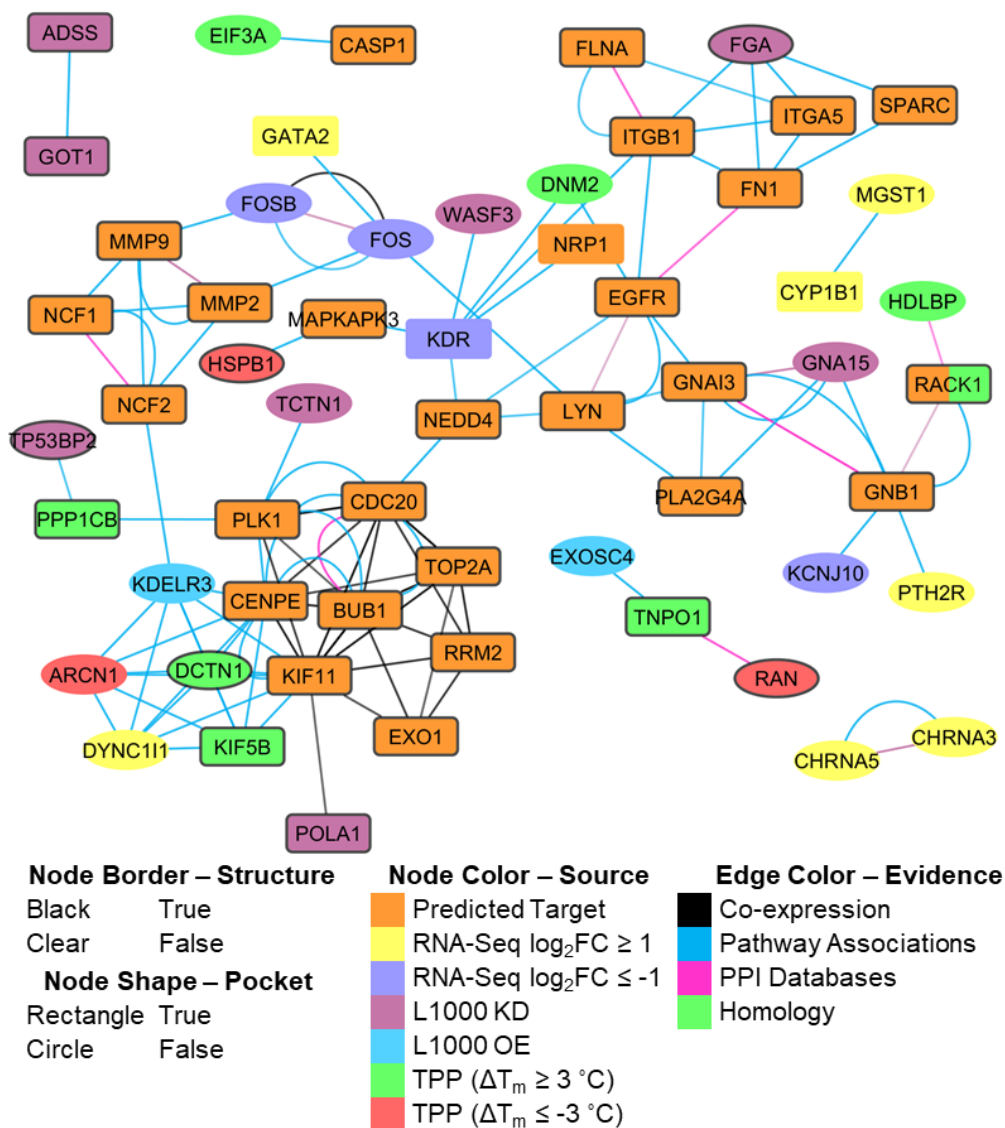


Figure 6.12. Comparison of proteins implicated by compound 1. The SVR-KB predicted targets (orange), differentially expressed genes (overexpressed: yellow, underexpressed: blue-purple), most similar perturbagens from L1000 analysis of differentially expressed RNA-seq genes (KD: purple, OE: blue), and proteins from thermal proteome profiling ($\Delta T_m \geq 3 \text{ }^\circ\text{C}$: green, $\Delta T_m \leq -3 \text{ }^\circ\text{C}$: red) are represented as nodes. Node borders are black if there is a solved human crystal structure available or uncolored otherwise. Nodes appear as rectangles if there is a druggable binding pocket on the associated protein or circles otherwise. Connections between nodes are built using STRING. Edges between connected proteins are filtered by confidence (STRING confidence: high, score ≥ 0.7) and colored based on the source of evidence (co-expression: black, database: blue, experimental: magenta). Unconnected nodes are omitted.

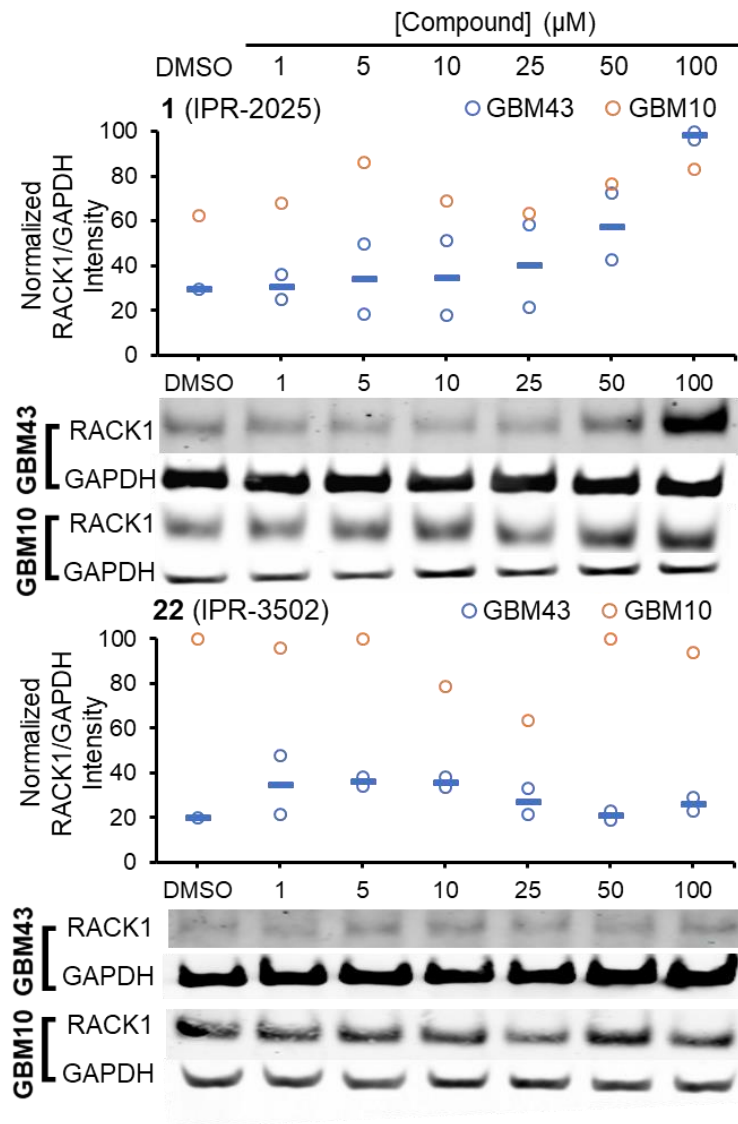


Figure 6.13. Validation of RACK1 as a target of compound **1**. Cellular thermal shift assay (CETSA) to determine direct binding of **1** (top) and inactive analog **22** (bottom) with RACK1. GAPDH act as negative controls. GBM43 cells were treated at various concentrations of compound and then heated to 45 °C. The normalized RACK1/GAPDH circles at each concentration represent biological replicates (n = 2), with a horizontal bar representing the mean intensity of the replicates.

6.2.7 Thermal Proteome Profiling to Identify Potential Targets of 1 (IPR-2025). Once a compound is identified in a phenotypic screen, the challenge is to identify its targets. Several methods have been used in the past for target identification, such as biochemical, genomic, or computational approaches [336-339]. Computational methods make use of statistical and machine learning methods to link a novel compound to existing compounds by identifying common features between the two. Previously developed algorithms have used similarities in compound scaffolds [340, 341], protein structure similarity [342], side effects [242], and bioactivities [343] to infer a compound's target. A more recent mass spectrometry-based method is thermal proteome profiling, which allows for systematic identification of a compound's direct targets by analyzing shifts in melting temperature of the targets in cells [344, 345]. As a compound binds to its protein target, the protein becomes more resistant to heat-induced unfolding and denaturation. This in turn increases the melting temperature of the protein.

Thermal proteome profiling was used to identify potential targets of **1**. GBM43 cells were treated with 10 μM of **1** at six temperatures between 37 and 60 $^{\circ}\text{C}$. Following heating, soluble proteins were extracted in phosphate-buffered saline (PBS) buffer, quantified and digested into peptides using trypsin. Digested peptides were analyzed by LC-MS/MS and quantified using label-free precursor-ion (MS1) intensity. The iBAQ values (intensities normalized by the size of the proteins) acquired after data analysis using MaxQuant [346] were used to fit thermal melting curves and to determine the shift in melting temperature between untreated and compound-treated proteins (ΔT_m). We acquired quantitative iBAQ data for over 1700 proteins across all temperatures at 1% FDR, of which the melting curves of 129 proteins were determined from the thermal profiling (**Fig. 6.11**). A cutoff of 3 $^{\circ}\text{C}$ was used to identify proteins with significant thermal shifts when treated with **1**. A total of 12 proteins were identified with $\Delta T_m \geq 3$ $^{\circ}\text{C}$ (**Table 6.3**). Interestingly, protein, *RACK1*, was both predicted to be a target of **1** (**Table 6.2**) as well as showed a significant thermal shift in TPP. In contrast, a set of 7 proteins were identified to be destabilized when exposed to **1**, resulting in a change in melting temperature $\Delta T_m \leq -3$ $^{\circ}\text{C}$. The largest observed destabilization was in *DSTN*, a component of the cytoskeleton.

6.2.8 Integrated Analysis of Computational, RNA-seq, and TPP Data for Potential Mechanisms of Action. To uncover the potential mechanism of action of **1**, proteins predicted to bind **1** by SVR-KB, and proteins implicated by compound-specific changes by RNA-seq profiling, L1000 comparisons, and thermal proteome profiling were integrated into a protein network (**Fig. 6.12**). Proteins implicated from each of the four sources were connected based on structural and non-structural evidence using the STRING database[347]. STRING incorporates protein-protein interactions from both direct physical interactions as well as indirect functional associations.

Interactions come from a variety of sources, including experimental interactions from protein-protein databases, pathway knowledge from manually curated databases, co-expression studies, and homology. Multiple interconnected modules are formed in the protein subnetwork. The largest module features genes associated with the cell cycle (*PPP1CB*, *PLK1*, *EXO1*, and *CDC20*) and metabolism of compounds with nucleobases (*KIF11*, *EXO1*, *TOP2A*, and *CENPE*). Within this cluster, several targets were predicted to bind to **1**, including *EXO1*, *TOP2A*, *CENPE*, and *KIF11*. Three proteins in this cluster showed positive thermal shifts, *DCTN1*, *KIF5B*, and *PPP1CB*, as well as a negative thermal shift in *ARCNI*. Both *DCTN1* and *KIF5B* act as motor proteins: *DCTN1* is a subunit of dynactin, which binds to dynein and acts as cytoskeletal motors in cellular transport[348], while *KIF5B* is a motor protein involved in mitosis and meiosis, and acts as a catalytic subunit of the tumor suppressors *NF1* and *NF2*[349]. Interestingly, upregulation of the dynactin mediator *DYNC1H1* is observed in the RNA-seq analysis. A second interconnected module is formed by the G protein β -subunit and β -subunit-like proteins *GNBI* and *RACK1*, respectively. The prediction of *RACK1* as a direct target of **1** is supported by direct evidence from the thermal proteome profiling, as well as a similar stability shift in the RNA-binding protein *HDLBP*. Similarly, upregulation of the GPCR *PTH2R* and downregulation of the potassium channel *KCNJ10*, which belong to similar regulation pathways as *GNBI*, was also observed in the analysis of RNA-seq derived from treated versus non-treated GBM spheroids.

To confirm direct binding to *RACK1*, the effects of **1** and the inactive analog **22** were examined in a concentration-dependent manner with an antibody-based cellular thermal shift assay (CETSA) (**Fig. 6.13**). Similar to thermal proteome profiling, GBM43 and GBM10 cells are treated with varying concentrations of the compound and then heated to 45 °C. Direct binding of the compound to the protein will result in protein stabilization and an increase in melting temperature. In GBM43, there is an increase in *RACK1* abundance at 50 and 100 μ M in cells treated with **1**, suggesting that *RACK1* is among the targets of **1**. To rule out non-specific binding of the compound, the assay was also performed on *GAPDH*, an enzyme involved in glycolysis, resulting in no difference in protein abundance with increase concentration of compound. When either *RACK1* or *GAPDH* were treated with the inactive analog **22**, the concentration of the compound did not affect protein abundance compared to the DMSO control.

Table 6.4. Predicted protein targets of **29** (IPR-196).

Symbol	Name	Family	Protein	PPI Network		TCGA GBM	
				Degree	Betweenness Centrality	Fold Change	Mutations
<i>ATP Binding Site</i>							
DDX39A	ATP-dependent RNA helicase DDX39A	DEAD box helicase		5	4.80E-04	1.8	5
<i>Protein-Protein Interaction Interface</i>							
PNP	Purine nucleoside phosphorylase	PNP/MTAP phosphorylase		2	0.00E+00	1.4	6
<i>Beta-Propeller</i>							
ITGA5	Integrin alpha-5	Integrin alpha chain		2	1.50E-05	3	11
<i>Phospholipid Binding Site</i>							
NR5A2	Nuclear receptor subfamily 5 group A member 2	Nuclear hormone receptor		6	3.10E-05	4.4	6
<i>Allosteric Site near Protein-Protein Interaction Interface</i>							
PYGL	Glycogen phosphorylase, liver form	Glycogen phosphorylase		3	0.00E+00	3.2	19
<i>Other Allosteric Sites</i>							
GLA	Alpha-galactosidase A	Glycosyl Hydrolase		1	0.00E+00	1.6	10
GUSB	Beta-glucuronidase	Glycosyl hydrolase 2		2	0.00E+00	2	7
ACE	Angiotensin-converting enzyme	Peptidase M2		1	0.00E+00	1.9	19
APOBEC3C	DNA dC->dU-editing enzyme APOBEC-3C	Cytidine and deoxycytidylate deaminase		3	4.50E-09	3.4	7
BCHE	Cholinesterase	Type-B carboxylesterase/lipase		2	0.00E+00	2.2	9
ITGB3	Integrin beta-3	Integrin beta chain		8	3.50E-04	3	16
MMP9	Matrix metalloproteinase-9	Peptidase M10A		4	1.60E-08	7.3	8
NRP1	Neuropilin-1	Neuropilin		2	4.10E-07	1.8	18
PLA2G4A	Cytosolic phospholipase A2			1	0.00E+00	1.6	7
RRM2	Ribonucleoside-diphosphate reductase subunit M2	Ribonucleoside diphosphate reductase		2	2.60E-06	7.2	5
TOP2A	DNA topoisomerase 2-alpha	Type II topoisomerase		3	2.50E-05	7.2	10

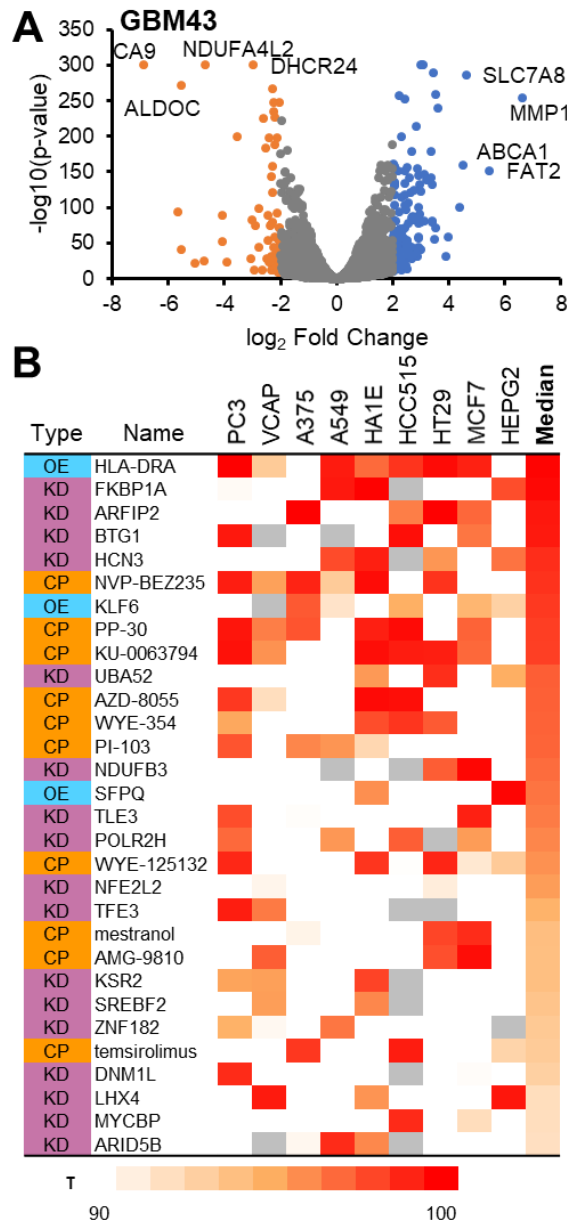


Figure 6.14. RNA-Seq of **29** (IPR-196) treated GBM43 cells. **(A)** Volcano plot of fold change versus log-transformed significance of GBM43 cells treated with 10 μ M of **29** versus control. Differential-expressed genes are identified using a \log_2 fold change cutoff of 2. **(B)** The most similar 30 perturbagens to the gene signature of treated GBM43 identified using the L1000 platform. Perturbagens are rank-ordered using the median τ statistic across nine cell lines, and are classified into gene knockdown (KD), overexpression (OE), and compounds (CP).

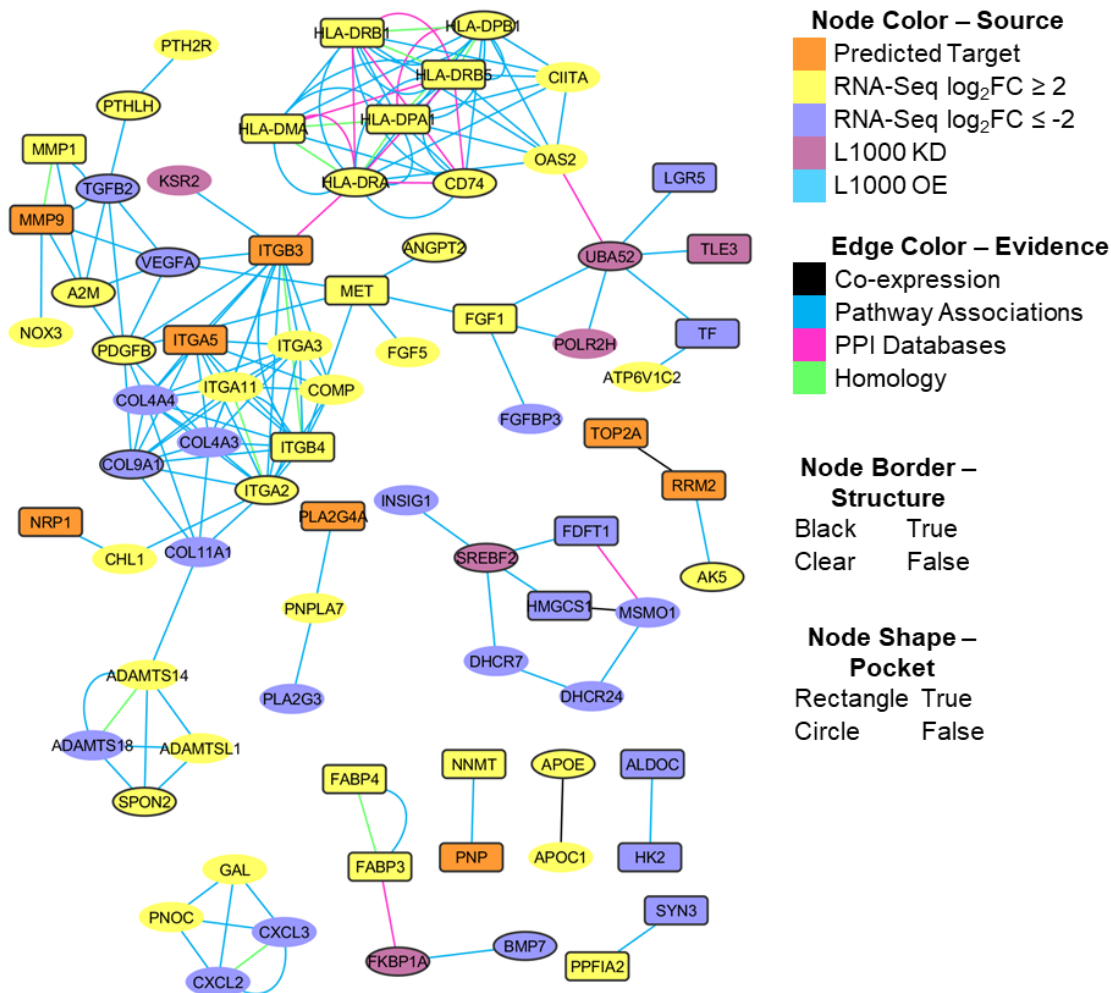


Figure 6.15. Comparison of proteins implicated by compound **29**. Comparison of proteins implicated by **29**. The SVR-KB predicted targets (orange), most similar perturbagens from L1000 analysis of differentially expressed RNA-seq genes (KD: purple, OE: blue), and differentially expressed genes (overexpressed: yellow, underexpressed: blue-purple) are represented as nodes. Node borders are black if there is a solved human crystal structure available or uncolored otherwise. Nodes appear as rectangles if there is a druggable binding pocket on the associated protein or circles otherwise. Connections between nodes are built using STRING. Edges between connected proteins are filtered by confidence (STRING confidence: high, score ≥ 0.7) and colored based on the source of evidence (co-expression: black, database: blue, experimental: magenta). Unconnected nodes are omitted.

6.2.9 Structural Analysis and RNA Sequencing to Uncover Compound 29 Mechanism of Action. The binding modes of **29** to the targets predicted by SVR-KB were further examined (Table 6.4). Like **1**, most predicted binding sites of **29** occur at allosteric binding pockets. However, the predicted binding sites do not share similar protein functions. Compound **29** binds to only two sites that are known to be critical to protein function: the ATP binding site of *DDX39A* and the phospholipid binding site of *NR5A2*. The RNA helicase *DDX39A* alters RNA for transcription, splicing, and editing that is driven by ATPase activity [350, 351]. Similarly, the nuclear receptor *NR5A2* (*LRH-1*) regulates bile-acid homeostasis and cholesterol transport, and features a hydrophobic ligand-binding pocket that normally binds phospholipids that mediates coactivator interaction and transcriptional activity [352]. The compound is also predicted to bind in the cavity formed by the FG-GAP repeat beta-propeller structure on the integrin *ITGA5*. It is also predicted to bind at homotrimer interface of *PNP* and at an allosteric pocket adjacent to the homodimer interface of *PYGL*.

RNA sequencing of GBM43 treated with **29** was carried out at 10 μ M to explore potential mechanism of action of the compound (Fig. 6.14A). Differential expression analysis was carried out and revealed a larger number of DEGs compared to **1**. Considering the large number of genes, a more stringent fold change cut-off was used for differential expression ($p < 0.001$, FDR < 0.01 , $|\log_2FC| > 2$). In total, 134 overexpressed and 65 underexpressed genes were identified. Gene overrepresentation analysis [324] of the overexpressed genes revealed two significant biological processes terms. The first term involves genes associated with cell-matrix adhesion (fold enrichment = 12.8, FDR < 0.04) and its parent term cell adhesion (fold enrichment = 4.1, FDR < 0.03), such as integrins, *FGL2*, *NTN4*, and *ANGPT2*. The second term is cellular defense response (fold enrichment = 9.3, FDR < 0.01), which involves a set of genes in the human leukocyte antigen (HLA) system. No significant GO processes were identified among the set of underexpressed genes. Comparison of the differentially expressed genes with those with known causal mutations in cancer [325] include the tumor suppressing transcription factor *KLF4* and oncogenes *MET* and *MYCL*. Similarly, *CD74* and *CIITA* are associated with MHC-II immune response, while the lncRNA *MALAT1* is associated with chemoresistance to temozolomide [353].

The L1000 platform [328] was used to identify known perturbations with the most similar expression signatures as **29** (Fig. 6.14B). The most similar pattern is overexpression of a gene in the HLA system, which is consistent with overrepresentation of genes involved in cellular defense response. The top knockdown signatures are genes involved in immunoregulation (e.g. *FKBP1A* and *TFE3*) and cell differentiation (e.g. *BTG1*, *UBA52*, and *TLE3*). Among the top compound signatures, the majority are known *mTOR* and *PI3K* kinase inhibitors. The sets of genes predicted

by SVR-KB, RNA sequencing, and L1000 platform were integrated using STRING[347] to uncover potential mechanisms of **29** (Fig. 6.15). Proteins implicated from each of the three sources were connected based on both structural and non-structural evidence. The gene signatures identified by L1000 are not directly associated with the set of SVR-KB genes. One exception is the connection between *ITGB3* and *KSR2* from SVR-KB predicted and L1000 knockdown, respectively, which are both involved in *MAP2K/MAPK* activation.

The largest interconnected module in the subnetwork is formed by upregulation of integrins (e.g. *ITGA2*, *ITGA3*, *ITGA4*, *ITGA11*, and *ITGB3*) and downregulation of collagens (e.g. *COL4A3*, *COL4A4*, *COL9A1*, *COL11A1*), which is associated with the SVR-KB predicted targets *ITGA5* and *ITGB3*. Similarly, our predicted target *MMP9* is associated with the observed upregulation of *NOX3*, *A2M*, and *MMP1* and downregulation of *VEGFA* and *TGFB2* in the RNA-seq. The gene signature knockdown of *SREBF2*, a transcription factor associated with cholesterol homeostasis, is consistent with downregulation of a variety of cholesterol biosynthesis and regulation, including *FDFT1*, *INSIG1*, and *HMGCS1*, but is not associated with any of our predicted targets. Treatment with **29** resulted in up-regulation of a collection of genes associated with MHC-II immune response, including multiple HLAs, *CD74*, and *CIITA*.

6.3 DISCUSSION

Tumors such as GBM exhibit multiple phenotypes that include uncontrolled growth, angiogenesis, invasion, and immune evasion. These phenotypes are driven by perturbations in genes and their protein products working in concert across multiple signaling pathways. The multiple targets involved in promoting tumor growth and metastasis pose a major challenge for the development of small-molecule therapeutic agents to treat these tumors. To date, the most common strategy in cancer drug discovery is to develop small molecules that modulate the function of a single target. This approach has not yielded therapeutic agents that are efficacious for complex tumors such as GBM. At the other extreme, phenotypic screening has been used to uncover novel anti-cancer agents [282]. This strategy has led to several approved drugs, including eribulin in breast cancer, nelarabine in T-cell lymphoblastic leukemia and lymphoma, and trametinib in metastatic melanoma [354]. However, the limited diversity of chemical libraries, the use of immortalized cell lines, and the reliance on two-dimensional cellular assays has mostly led to cytotoxic cell cycle inhibitors that have not yielded efficacy in patients [355].

We believe that a data-driven approach that combines genomic, structural, and protein-protein interaction data to enrich chemical libraries using computational docking has the potential to overcome the limitations of phenotypic screening for cancer drug discovery. Here, in a proof-of-

concept study, we propose a screening strategy that takes advantage and integrates vast orthogonal datasets including (i) tumor genomic data from patient-derived GBM samples available at TCGA, (ii) three-dimensional structures of human proteins that enable the identification of druggable pockets in proteins implicated in GBM, and (iii) the large number of cellular protein-protein interactions that have been mapped over the past decade using yeast-two-hybrid and other methods. Specifically, our approach initially utilizes expression data from TCGA to select all genes that are overexpressed and mutated in GBM. The PDB is subsequently mined to retrieve available protein structures that are encoded by these genes. Druggable pockets within these structures are identified and used for structure-based screening to identify potential small-molecule inhibitors. Proteins that are known to be involved in protein-protein interactions and possess druggable pockets are used for structure-based docking of an in-house library of 9,000 compounds. The resulting protein-compound complexes were ranked to select 50 small molecules that bind to the highest number of GBM-specific proteins. In another strategy, we selected 50 compounds that bind to 20 predicted GBM-specific targets or less. The resulting 100 compounds were tested in a cell viability assay utilizing patient-derived GBM cells grown in a physiologically relevant three-dimensional format. Our strategy has a significant advantage over standard phenotypic screening, namely that large libraries containing thousands of compounds can be enriched to select a small collection of candidates that can be tested in more sophisticated assays and multiple phenotypes. Two assays are used, spheroid growth and tube formation assay in Matrigel. We also strictly use patient-derived low-passage cell lines to overcome the limitations of existing phenotypic screens that are done on established cell lines. In addition to using GBM cancer cell lines, we also employ low passage pancreatic ductal adenocarcinoma cell lines as well as normal non-transformed cells such as CD34⁺ progenitor cells and astrocytes.

Compound collections were initially screened at a single concentration in duplicate using GBM43 cells grown in 3D spheroids. In the first set of compounds that were predicted to maximally bind to GBM-specific targets, one compound, namely **1** (IPR-2025), showed substantial inhibition of cell viability. For the second set of compounds that were predicted to bind to 20 or less GBM-specific targets, three hits emerged, **28** (ALDH-22), **29** (IPR-196), and **30** (IPR-1964). All three compounds inhibited GBM spheroid viability by 80% or more at 25 μ M. A concentration-dependence study revealed that **1** (IPR-2025) showed the highest potency in all three patient-derived GBM cell lines with IC_{50} s < 4 μ M, while the remaining compounds had IC_{50} s that are 10 μ M or greater. All four were tested for their effect on tube formation in Matrigel, an indication of their potential usefulness in blocking angiogenesis and a key feature of GBM tumors, with sub-micromolar IC_{50} s. Two compounds, **1** and **29**, were tested in both normal non-transformed human

CD34⁺ progenitor cells and primary brain astrocytes. Significantly, **1** had no effect on colony formation of CD34⁺ cells or astrocyte cell viability suggesting that there is a therapeutic window. Compound **29** also demonstrated negligible inhibition of CD34⁺ cell viability, but some cytotoxicity is detected at 30 μ M or higher for astrocytes. The toxicity may be due to the lack of solubility of the compound at these higher concentrations since the compound shows substantial inhibition of GBM cell viability despite the lack of toxicity in astrocytes and CD34⁺ cells. To put these activities in perspective, standard-of-care temozolomide inhibits GBM43 growth with an approximate IC₅₀ of 250 μ M consistent with the high therapeutic doses used in GBM.

RNA-seq was initially employed to uncover a potential mechanism of action for **1** and **29** in GBM43 cell lines. Both compounds led to up-regulation and down-regulation of genes in cancer cells. Compound **1** exhibited remarkable selectivity as relatively fewer genes were affected by this compound compared to **29**. The L1000 platform was employed to identify gene knockdowns that led to a similar effect on gene expression. However, L1000 utilizes adherent cell cultures and established cell lines in cancers other than GBM. In contrast, GBM43 is a low-passage patient-derived cell line grown as a spheroid model and is expected to have a different underlying gene expression profile. Interestingly, the expression profile of **1** was most similar to the knockdown of G protein *GNA15*, which belongs to the same family as the predicted target *GNBI* and structural homolog *RACK1*.

Considering the GBM-specific activity of **1** (IPR-2025) and its lack of cytotoxicity in normal non-transformed cell lines, thermal proteome profiling was used to identify potential targets of this compound. TPP uncovered up to 11 potential targets of **1**, including *RACK1* and *KIF5B*. It is interesting that among them, one target (*RACK1*) was also among the targets of **1** that emerged from the structure-based docking and ranking analysis. Analysis of the predicted binding mode reveals that the compound binds within the central tunnel of the β -propeller structure. The scaffolding protein *RACK1* has been shown to be upregulated during angiogenesis[356], and *RACK1* in complex with *VIM* was shown to regulate angiogenesis by modulating *PTK2/FAK1*[357], providing further evidence that *RACK1* could be a potential target of **1**. Follow-up CETSA analysis seems to support the fact that **1** may bind to *RACK1* directly in cell culture. It is important to note that the TPP list is likely not a comprehensive list of all direct targets of **1**, and it is also likely that the 20 top targets identified in TPP may not be all targets of **1**, as the compound was designed to bind and modulate multiple targets.

In summary, we developed a multi-target approach that integrates cancer genomics with the druggable protein interactome to identify therapeutic candidates of GBM. Either strategy in selecting compounds that were predicted to either target the highest number of GBM-specific

proteins or to a lower threshold of 20 or less GBM-specific proteins yielded candidate compounds. We pursued two compounds in RNA-seq, one from the first approach (compound **1**) and another from the second approach (compound **29**). Thermal proteome profiling of **1** revealed possible targets, including the computational predicted target *RACK1*. The ability of **1** to selectively target GBM phenotypes without affecting normal cell viability makes it suitable as a lead compound to uncover new targets in GBM and to develop potential therapeutic agents with multi-targeting properties.

6.4 MATERIALS AND METHODS

6.4.1 TCGA GBM Gene Expression and Somatic Mutations. RNA-sequencing and mutation data from GBM were identified from The Cancer Genome Atlas (TCGA) project [12, 217] and collected from the National Cancer Institute's Genomic Data Commons (GDC) data portal [358]. As part of the GDC pipeline, 169 tumors and 5 normal samples were previously aligned against the GRCh38 genome assembly using STAR 2-pass [359] and quantified using HTSeq [360]. Level 3 HTSeq fragment counts were collected from all 174 samples and used these as read counts for differential expression analysis. Differentially expressed genes were identified using the edgeR [266] package and the quasi-likelihood F-test pipeline [361] for hypothesis testing. A counts-per-million (CPM) threshold corresponding to a library count of 10 ($\text{CPM} \approx 0.15$) was used to filter out genes with low reads. Overexpressed genes were defined as those with $p < 0.001$, $\text{FDR} < 0.01$, and \log_2 fold change ($\log_2\text{FC}$) > 1 . Ensembl Gene identifiers from the differential expression analysis were mapped to UniProt [119] identifiers using BioMart [362]. Open-access mutation annotation format (MAF) files for GBM were also collected from the GDC data portal. The GDC pipeline used the MuTect2 [363] pipeline for mutation calling. As part of the GDC's workflow, low quality and potentially germline variants were removed. Somatic mutations were mapped to their corresponding proteins using the UniProt ID.

6.4.2 Protein-Protein Interaction Network. Rolland and associates [308] recently constructed two large-scale protein-protein interaction networks. The first is based on literature curation and benchmarking of seven protein-protein interaction databases, producing a network of approximately 5500 proteins and 12000 interactions. The second is based on systematic experimental testing of pairwise combinations of approximately 13000 genes and describes a network of approximately 4200 proteins and 14000 interactions. Entrez genes were mapped to UniProt identifiers using UniProt's mapping tool (<http://www.uniprot.org/mapping/>) and these two networks were combined together using the NetworkX [273] module in Python. The resulting

protein-protein interaction network consists of approximately 27000 interactions across 8000 proteins and was visualized using Cytoscape [364] (v3.3.0).

6.4.3 Druggable Binding Sites. In a previous work, a set of druggable binding pockets were identified in the human proteome and explored in the context of functional importance, signaling pathways, and the human interactome [307]. In brief, a set of 4124 proteins were identified in the human proteome that have been solved with crystallography. Druggable binding pockets were found on these proteins using the SiteMap [221, 257] module in Schrödinger (Schrödinger LLC, New York, NY). Up to 10 binding sites were mapped for each structure. Each binding site was evaluated by its ability to bind a ligand (SiteScore) and its druggability (DrugScore). Only binding sites with SiteScore and DrugScore above 0.8 were kept. In total, 5498 binding sites on 2607 proteins were found. Each binding site identified by SiteMap was visually inspected and manually annotated to determine its functional role in the protein. If an active site residue was in contact with the SiteMap spheres, or if a catalytic molecule or inhibitor occupied the space of the spheres, the binding site was labeled ‘enzyme’ (ENZ). If the binding site was at a protein-protein interaction (PPI) interface on the original structure or on any of the aligned structures, the binding site was labeled ‘PPI’. Otherwise, if the binding site was neither enzyme nor part of the interaction interface, it was labeled ‘other’ (OTH).

6.4.4 Virtual Screening Against GBM Network. The set of druggable binding sites on proteins which were (i) overexpressed in GBM, (ii) featured somatic mutations in GBM, and (iii) were part of the PPI interactome were selected. Specifically, we focused on binding sites that were classified as either PPI or OTH and omitted those with solely ENZ classifications. In total, 316 binding sites were identified. The set of binding pockets were docked against an internal chemical library of small molecules. These compounds are from previous screening campaigns primarily focused on targeting tight protein-protein interactions. Compounds possessing pan-assay interference compound (PAINS) [170] or rapid elimination of swill (REOS) [171] moieties were discarded. This resulted in a compound library of 9075 small molecules and their enantiomers. Compounds were docked against 316 binding sites on proteins implicated in GBM using AutoDock Vina [172] (v1.1.2). The average coordinates of the SiteMap spheres were used to identify the centroid of the binding site for docking. Each binding site was represented as a box with dimensions of $21 \text{ \AA} \times 21 \text{ \AA} \times 21 \text{ \AA}$. All other parameters were set to default values. Docked poses were rescored using the previously developed Support Vector Machine-Knowledge Based (SVR-KB) [309] scoring function. In this scoring function, knowledge-based pair-potentials of co-crystal complexes are used to train a regression model to predict the binding affinity of docked complexes. To rank-order compounds, the number of pockets with binding affinities better than a given SVR-KB cutoff

for that compound were counted. Here, an SVR-KB cutoff corresponding to a predicted binding affinity of 10 nM was used. Compounds were rank-ordered using the number of binding sites predicted to bind to the compound with an affinity better than the SVR-KB cutoff. The top 154 compounds were retrieved and clustered using the Canvas module in Schrödinger. A hashed binary fingerprint corresponding to atom triplets of Daylight invariant atom types were generated for the selected compounds. Compounds were hierarchically clustered using their atom triplet fingerprints and average linkage clustering to 50 clusters. The Tanimoto similarity between a pair of fingerprints was used as the distance metric. Compounds corresponding to the centroid of each cluster were selected for experimental validation.

6.4.5 Cell Culture. GBM43, GBM10, and SJ-GBM2 cells were cultured in DMEM medium with glutamine (Cellgro, Manassas, VA) supplemented with 10% FBS and 1% penicillin/streptomycin in 5% CO₂ at 37 °C.

6.4.6 Three-Dimensional Culture Models. The GBM43 cell line was generated as previously described [314]. The GBM43 xenograft tissue was a kind gift from Dr. Jann Sarkaria (Mayo Clinic, Rochester, IN), and tumors were expanded by passage in the flank of NOD/SCID γ^{null} mice. To generate GBM43 cell lines, tumors were harvested, disaggregated, and maintained in 2.5% FBS for 14 days on Matrigel-coated plates (BD Biosciences) to remove murine fibroblasts. In-vitro GBM43 cell lines were propagated in DMEM with 10% FBS for no more than 7 passages. Cell line identity was confirmed by DNA fingerprint analysis (IDEXX BioResearch) for species and baseline short-tandem repeat analysis testing. GBM43 spheroids were generated by plating early-passage cells at 2.5×10^4 cells per well in 96-well ultralow attachment plates (Corning Inc.) in DMEM/F12 (1:1; GIBCO) supplemented with 2% B27 supplement (GIBCO), 20 ng/mL epidermal growth factor (EGF), and 20 ng/mL fibroblast growth factor (FGF) (Peprotech) for 2 days. The spheroids were then treated with compounds and growth analyzed by Alamar blue staining at day 5 following compound exposure.

Astrocyte cell proliferation was determined by The CellTiter 96® AQueous Non-Radioactive Cell Proliferation Assay (MTS) (Promega) performed in 96-well plates. Cells were seeded at 8×10^3 /well in DMEM medium with 10% FBS and 1% 100X Penicillin/Streptomycin. Cell numbers were determined after 3 days of incubation with DMSO or compounds at indicated concentrations. 20 μ L of MTS solution was added to the well. After 1 h incubation at 37 °C, the absorbance was measured at 490 nm. Experiments were done in triplicate.

Half maximal inhibitory concentration (IC₅₀) of cell viability curves were determined using a four-parameter logistic model [365]. IC₅₀ values represent relative IC₅₀s, that is, the minimum

parameter of the model is the lowest cell viability observed. Cell viability curves were determined using the non-linear least-squares method as implemented in the SciPy package in Python.

6.4.7 Invasion Assay. The Matrigel based tube formation assay was performed as previously described [366]. Briefly, 50 μ L Matrigel (Corning, Corning, NY) was allowed to solidify in a 96 well black, clear bottom plate at 37 °C for 20 min. Primary brain microvascular endothelial cells (BMECs; Cell Systems, Seattle, WA) were added to the solid Matrigel at 15,000 cells per well in 100 μ L endothelial growth medium-2 (EGM-2, Lonza, Walkersville, MD) and dosed with appropriate concentrations of compound with 1 μ L DMSO per well. Tube formation was observed every 2 hours by brightfield microscopy and images were taken after 8 hours of tube formation. Six images per treatment were analyzed with the Angiogenesis Analyzer plugin for ImageJ [367], and BMEC total tubule length for treated cells was normalized to DMSO-treated samples. Statistical analysis using one-way ANOVA with Dunnett's *post hoc* test to compare treatments with DMSO control was completed using GraphPad Prism (v7.0, GraphPad Software, La Jolla, CA).

6.4.8 RNA-Seq of Compound-Treated Cells. GBM43 cells were treated with 10 μ M **1** (IPR-2025) and **29** (IPR-0196). RNA-seq analysis was conducted in triplicates for the untreated control, **1**, and **29**. After compound treatment, GBM43 cells were collected and rinsed with 1 \times PBS. The pellet was first homogenized in RLT lysis buffer plus β -mercaptoethanol and total RNA was extracted with RNeasy Mini Kit (Qiagen) in combination with QIAshredder (Qiagen). The RNA quality was assessed by A260/A280 ratio (NanoDrop) and stored at -80 °C.

Total RNA was evaluated for its quantity and quality using the Agilent Bioanalyzer 2100 system. For RNA quality, a RIN number of 7 or higher was desired. A total of 500 ng RNA was used. The cDNA library was prepared through mRNA purification and enrichment, RNA fragmentation, cDNA synthesis, ligation of index adaptors, and amplification, following the TruSeq Stranded mRNA Sample Preparation Guide, RS-122-9004DOC, Part# 15031047 Rev. E (Illumina, Inc.). Each resulting indexed library was quantified, and its quality accessed by Qubit and Agilent Bioanalyzer, and multiple libraries pooled in equal molarity. Five μ L of 2 nM pooled libraries per lane was then denatured, neutralized, and applied to the cBot for flow cell deposition and cluster amplification, before loading to HiSeq 4000 for sequencing (Illumina, Inc.).

The RNA-seq aligner from STAR [359] (v2.5) was used to map RNA-seq reads to the human reference genome (hg38), with the following parameter: '--outSAMmapqUnique 60'. Uniquely mapped sequencing reads were assigned to genes using featureCounts [368] (from subread v1.5.1) with the following parameters: '-s 2 -Q 10'. The genes were filtered for further analysis if their count per million (CPM) of reads was less than 0.5 in more than 4 samples. The

method of trimmed mean of M values (TMM) was adopted for gene expression normalization across all samples, followed by differential expression (DE) analysis between different conditions using edgeR [266, 369] (v3.20.8).

Differentially expressed genes (DEG) were determined using p-value and false discovery rate (FDR) cutoffs of 0.001 and 0.01, respectively. A fold change ($|\log_2 FC|$) cutoff of 1 and 2 were used for **1** and **29**, respectively. The functional analysis was performed on overexpressed and underexpressed DEGs separately with a cutoff of $FDR < 0.05$ to identify significantly overrepresented Gene Ontology (GO) and/or KEGG pathways using DAVID [370, 371] and PANTHER [324].

6.4.9 Thermal Proteome Profiling. GBM43 cells were harvested by centrifugation at 500g and washed twice with PBS and protease inhibitor cocktail. Cells were then treated with 10 μM **1** or DMSO (as control) for 24 h. Treated cells were pelleted again and re-suspended in ice-cold PBS. This was repeated twice. Pellets were then exposed for 3 min to the following temperatures: (i) 37 °C, (ii) 40 °C, (iii) 45 °C, (iv) 50 °C, (v) 55 °C, and (vi) 60 °C. After a freeze-thaw treatment, cells were pelleted again, and the supernatant was snap frozen in liquid nitrogen.

After thaw, protein concentration in each sample was determined by bicinchoninic acid (BCA) assay with BSA as a standard. About 50 μg protein from each sample was denatured by adding 50 μL of 8M urea, reduced by incubating with 10 mM dithiothreitol (DTT) at 50 °C for 45 min, and cysteine alkylated with 20 mM iodoacetamide (IAA) in the dark for 45 min at room temperature, followed by incubation with 5 mM DTT for 20 min at 37 °C to scavenge residual IAA. Proteins were digested using sequencing grade trypsin and Lys-C mix from Promega at a 1:25 (w/w) enzyme-to-protein ratio at 37 °C overnight. The digested peptides were cleaned using C18 silica micro spin columns (The Nest Group Inc.) and peptides were eluted using 80% acetonitrile containing 0.1% formic acid (FA). The samples were vacuum dried and re-suspended in 3% acetonitrile and 0.1% formic acid. The peptide concentration was determined by BCA assay with BSA as a standard. Peptide concentration was adjusted to 0.2 $\mu\text{g}/\mu\text{L}$, soluble and insoluble samples were mixed together and 5 μL was used for LC-MS/MS analysis as described below.

Samples were analyzed by reverse-phase LC-ESI-MS/MS system using the Dionex UltiMate 3000 RSLCnano System coupled to the Q Exactive High Field (HF) Hybrid Quadrupole-Orbitrap MS and a Nano-electrospray Flex ion source (Thermo Fisher Scientific). Peptides were loaded onto a trap column (300 μm ID \times 5 mm) packed with 5 μm 100 Å PepMap C18 medium, and then separated on a reverse phase column (15-cm long \times 75 μm ID) packed with 3 μm 100 Å PepMap C18 silica (Thermo Fisher Scientific). All the MS measurements were performed in the positive ion mode, using 120 min LC gradient as previously described [372]. The mass

spectrometer was operated using standard data-dependent mode. MS data were acquired with a Top20 data-dependent MS/MS scan method. The full scan MS spectra were collected in the 400-1,600 m/z range with a maximum injection time of 100 milliseconds, a resolution of 120,000 at 200 m/z, spray voltage of 2 and AGC target of 3×10^6 . Fragmentation of precursor ions was performed by high-energy C-trap dissociation (HCD) with the normalized collision energy of 27 eV. MS/MS scans were acquired at a resolution of 15,000 at m/z 200 with an ion-target value of 1×10^5 and a maximum injection of 20 milliseconds. The dynamic exclusion was set at 15 s to avoid repeated scanning of identical peptides. Instrument was calibrated at the start of each batch run and then in every 72 hours using calibration mix solution (Thermo Scientific). The performance of the instrument was also evaluated routinely using complex *E. coli* digest purchased from Sigma.

Raw LC-MS/MS data were analyzed using MaxQuant [346] (v1.6.0.16) against the UniProt human protein database containing 40,707 proteins. The database search was performed with the precursor mass tolerance set to 10 ppm and MS/MS fragment ions tolerance was set to 20 ppm. Database search was performed with enzyme specificity for trypsin and LysC, allowing up to two missed cleavages. Oxidation of methionine and N-terminal acetylation were defined as a variable modification, and carbamidomethylation of cysteine was defined as a fixed modification for database searches. The ‘unique plus razor peptides’ were used for peptide quantitation. The false discovery rate (FDR) of both peptides and proteins identification was set at 0.01. Data were searched with ‘match between runs’ option. In the case of identified peptides that are shared between two proteins, these were combined and reported as one protein group. Proteins matching to the reverse database were filtered out.

Protein abundances were quantified using iBAQ [373] and transformed into relative concentrations relative to the lowest temperature (37 °C). Protein fold changes were normalized and melting curves were fitted following the protocol described by Savitski [344], Franken [345], and co-workers. The previously described normalization protocol [344] was adapted to fit the range of temperatures in our experiments. Melting curves were selected from both the vehicle and treatment groups where the fold change at 55 °C was between 0.4 and 0.6 and the fold change at 60 °C was less than 0.4. Normalization was then applied as previously described [344]. Following normalization, melting curves were fitted into the sigmoidal function [345]:

$$f(T) = \frac{1 - p}{1 + e^{-\left(\frac{a}{T} - b\right)}} + p$$

where T is the temperature, a and b are constants, and p is the plateau as $T \rightarrow \infty$. The melting point T_m is found using:

$$T_m = \frac{a}{b - \ln\left(\frac{1-y}{y-p}\right)}$$

at $f(T_m) = y = 0.5$. The slope of the melting curve $slope = f'(T_{inflection})$ is found at the inflection point $T_{inflection}$ where $f''(T_{inflection}) = 0$. Melting curves were determined using non-linear least-squares method implemented in the SciPy package in Python. Proteins with poorly fitted curves were excluded if the parameters of the curve fit any of the following criteria: (i) vehicle or treatment curve with $R^2 < 0.8$, (ii) vehicle $plateau > 0.3$, (iii) the number of points between $(plateau + 0.1, 0.9) < 1$, and (iv) vehicle or treatment $slope > -0.06$. In total, 129 proteins were identified that passed all quality control filters. The shift in melting point ΔT_m was determined as:

$$\Delta T_m = T_m^{\text{treatment}} - T_m^{\text{vehicle}}$$

A cutoff of $\Delta T_m \geq 3.0$ was used to identify proteins with significant thermal stability following compound treatment.

6.4.10 Gene Set Analysis. The STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database [347] incorporates protein-protein interactions from both direct physical interactions as well as indirect functional associations. STRING was used to integrate gene sets identified from protein target prediction using SVR-KB, differentially expressed genes identified from RNA sequencing, most similar gene signatures from L1000, and genes identified by thermal proteome profiling. A concatenated list of all HUGO gene symbols was used as input. Connections coming from text mining sources were excluded. Edges with confidence score < 0.7 (high confidence) and genes with no edges in the subnetwork were excluded. The resulting subnetworks were visualized using Cytoscape [364].

6.4.11 Cellular Thermal Shift Assay (CETSA). GBM43 and GBM10 cells were treated with **1** (IPR-2025) and **22** (IPR-3502) as previously described [374]. Cells were treated for 6 h in CO₂ incubator at 37 °C. The treated samples were harvested and washed twice in PBS and suspended in PBS supplemented with a protease inhibitor cocktail (Sigma-Aldrich). Each sample with 5×10^6 cells was heated at 45 °C for 3 min, then incubated for 3 min at room temperature, and frozen in liquid nitrogen. Samples were lysed with three freeze–thaw cycles using liquid nitrogen and a 25 °C water bath. Cell lysates were centrifuged at $100,000 \times g$ for 20 min at 4 °C to separate protein aggregates from soluble proteins. Supernatants were collected for western blot with RACK1 and GAPDH antibodies (Cell Signaling).

Chapter 7

SUMMARY

7.1 CONCLUSION

Protein-protein interactions (PPIs) control almost every aspect of normal cellular function. The phenotypes observed in diseases such as cancer is driven by perturbations in protein-protein interactions as a result of underlying driver mutations. Despite their importance in cell signaling, protein-protein interactions remain unviable targets in traditional drug discovery. This can be attributed to the large, flat, and featureless interaction interfaces that are typical of protein-protein interactions. However, recent advances have identified potential design strategies for targeting protein-protein interactions.

Understanding how existing compounds engage and mimic hot spots could help guide structure-based computational screening of chemical libraries for the discovery of small-molecule protein-protein interaction inhibitors. Thus, in Chapter 2, existing protein-protein interaction inhibitors were explored with respect to the native protein ligand. First, existing inhibitors were collected for which there is a crystal structure of both the protein-protein and protein-compound complex. Computational methods were used to identify hot spots at PPI interface using alanine scanning and per-residue energy decomposition as a means to explore engagement of compounds with hot spots on the protein receptor and overlap between compounds and ligand hot spots. It was found that in general, existing small-molecule inhibitors of protein-protein interactions do not engage and mimic hot spots on the protein receptor and ligand, respectively. Designing compounds that make better use of hot spots may represent a viable strategy for the discovery of more effective inhibitors of protein-protein interactions.

The goal of the second part of this work was to identify potential therapeutics for protein-protein interactions. A traditional drug discovery campaign usually starts with structure-based virtual-screening to identify lead compounds. This process involves molecular docking of a compound library to a binding pocket on a target of interest to predict the binding pose of each compound. Binding poses are then evaluated with the use of scoring functions. Traditional scoring functions are developed for traditional binding pockets at active sites on enzymes and are not able to capture compound engagement with either receptor or ligand hot spots. In Chapter 3, a scoring method was developed to evaluate compound engagement with receptor hot spots. The scoring function made use of pairwise interactions of a compound with individual residues on the protein receptor. This fingerprint method was used to rank-order compounds based on their ability to mimic the interactions seen in the native protein ligand. The fingerprint method, which captures

engagement with receptor hot spots, was combined with a pharmacophore modeling method, which captures mimicry of ligand hot spots. This led to small molecules with novel chemotypes that inhibited a tight protein-protein interaction with single-digit micromolar binding affinities, suggesting that mimicking the binding profile of the native ligand and the position of interface residues can be an effective strategy to enrich commercial libraries for small-molecule inhibitors of tight protein-protein interactions.

An additional challenge in structure-based virtual-screening is the selection of the compound library. Few studies have examined the compounds in existing commercial and specialized collections for PPI drug discovery. In Chapter 4, we explored how effectively small molecules in commercial library, a collection of diversity-oriented synthesis libraries, and a compound collection feature fragment-like compounds mimicked the positions of critical residues at tight protein-protein interactions interfaces. A combined docking and pharmacophore approach was used to measure the overlap between compounds and sidechains of interface residues of three distinct PPIs. Our results suggested that while small fragment-like conformationally-restricted compounds can provide good candidates for PPIs with well-defined pockets, PPIs with large, flat, and featureless surfaces may require larger, more complex compounds.

In contrast to the first and second part of the work that examined individual protein-protein interactions, the final part of this work explores targeting protein-protein interaction networks. Diseases such as cancer exhibit multiple phenotypes such as uncontrolled cell growth, evasion of programmed cell death, and tumor invasion and metastasis. This is a result of perturbations in multiple protein-protein interactions in the global PPI network. Thus, a potential therapeutic for a disease such as cancer will not have one specific target, but rather a collection of targets that collectively combat the disease phenotypes. In Chapter 5, we characterize the molecular landscapes of several cancers to identify a subset of genes and their protein products that are altered in these cancers. Specifically, we look for genes that are highly overexpressed and mutated with druggable binding pockets on their respective protein products. Many of these altered proteins do not have prototypical enzyme active sites traditional seen in single-target drug discovery efforts and can only be targeted through their protein-protein interactions. We explored the role of these genes in the context of cancer related signaling pathways and the protein-protein interaction network to identify putative candidates for drug discovery.

An alternative to target-based process in drug discovery is the use of phenotype-based screens. Compared to traditional single target-based approaches, phenotypic screening identifies compounds that affect a specific phenotype, for example, cell viability or angiogenesis. However, traditional phenotypic screening is plagued by the identification of non-specific compounds and

inaccuracies in the experimental assays. To allay these issues, we describe a data-driven rational approach to create tumor-specific chemogenomic libraries for phenotypic screening of glioblastoma multiforme in Chapter 6. The approach combines catalogs of differentially expressed molecular targets identified by tumor genomic profiles along with cellular protein-protein interaction data to select a collection of targets with druggable binding pockets. Among the active hit compounds from phenotypic screening was a compound that inhibited cell viability of GBM spheroids and blocked tube-formation of endothelial cells but had no effect on primary hematopoietic progenitor or astrocyte cell viability. The success of the compound suggests that this approach to create a tumor-specific chemogenomic library may hold promise for cancer.

Many of the chapters of this thesis have been adapted from published works. These include Chapters 2 [186], 3 [178], 4 [375], and 5 [307]. Permission for adapting these works are included in **Appendix A**.

7.2 SUGGESTED FUTURE WORK

The computational results and experimental validation demonstrate that tight protein-protein interactions remain difficult targets for structure-based virtual design efforts in drug discovery. There are additional opportunities to build upon the previous described projects to identify initial small molecule candidates of protein-protein interactions for diseases such as cancer.

The first portion of this work was focused on identifying protein-protein inhibitors of individual targets. Initially, a set of co-crystallized inhibitors was selected for five specific protein-protein interactions. This represents only a portion of the vast structure-activity relationship data present of potential inhibitors for a single interaction. In the work, we only identified compounds with high-quality experimental data and a co-crystallized pose. One suggestion would be to include additional compounds with close co-crystallized analogs, and then modeling the binding poses of these additional compounds. The addition of more compounds would allow for more detailed examination of critical interactions in these tight protein-protein interactions.

In the second work, which described a fingerprint approach to rank-order compounds based on similar engagement of receptor hot spots, further work could focus on the selection of residue for the individual fingerprints. In this approach, a strict cutoff was used to differentiate whether the protein receptor interacted with the protein ligand. An alternative strategy would be to compare bits in the fingerprint using the interaction energy of the native complex. However, this raises additional concerns, particularly in the relative difference in binding energy when using fingerprints from either energy decomposition or alanine scanning. We showed in the first project that there is high

correlation at individual residues between energy decomposition and alanine scanning, but there were also examples of residues where there were profound differences in the associated energies.

In the third work, we looked at how different chemical libraries mimicked ligand hot spots for specific protein-protein interactions. Future work in this area could potential focus on using additional chemical libraries. Different combinatorial libraries, for example DNA-encoded libraries, offer additional opportunities that cover smaller areas of chemical space more robustly could be used for specific protein-protein interactions. Similarly, there is increasing interest in the development of PPI-specific libraries, which offer, for example, specifically alpha-helix or beta-sheet mimetics.

The final part of this work explored identifying cancer-relevant targets at protein-protein interactions. Here, we identified high-quality human crystal structures associated with individual proteins. There are two areas to consider for future works, specifically, the use of (i) high resolution crystal structures from (ii) humans. We did not consider close homology models or crystal structures with high sequence identity but featured a homologous protein in another species. Often, there are crystal structures of protein domains which have near identical sequence to the associated human protein. For these cases, manually mutations of non-identical residues offer an opportunity to consider a greater proportion of available structures. Similarly, we used a clustering algorithm to filter redundant structures of the same sequence. While this dramatically decreases the number of structures to consider, it does not account for protein dynamics, which may lead to transient binding sites on these structures that may have been captured in an alternative conformation. Similarly, molecular dynamics simulations of these structures, while computational infeasible, may be an alternative strategy to identify druggable binding sites.

Finally, we focused on targeting binding pockets implicated in GBM. Here, we focused on targets that were (i) overexpressed, (ii) mutated, (iii) had known PPIs, and (iv) had druggable binding sites. Somatic mutations in cancer can be classified as either driver or passenger mutations. In this work, we did not differentiate between the two. Similarly, genes that are mutated in GBM are not necessarily also overexpressed. This greatly limits the set of potential targets that were considered for this work. Another follow-up of this work could also focus on the selection of compounds. Here, we selected compounds that were the most promiscuous towards the set of GBM targets. Fortunately, we discovered through experimental follow-up that our hit compound was select to GBM. However, this should be considered in the initial selection, i.e., identifying compounds that selectively bind to GBM targets but not to non-GBM targets. We did not consider it in this study, mainly due to the already limiting selection of targets that already met the previous criteria.

In summary, we have described a series of studies related to inhibiting tight protein-protein interactions. First, we looked how existing protein-protein interactions inhibitors disrupt their targets. Second, we develop methods to identify candidates for individual protein-protein interactions by selecting compounds that mimic hot spot residues via both a scoring method and by enriching chemical libraries. Third, we identify potential protein-protein interactions that are relevant in the context of cancer. Finally, we developed a multi-target approach that integrates cancer genomics with the druggable protein interactome to identify therapeutic candidates of GBM.

APPENDICES

APPENDIX A. REPRINT PERMISSIONS FOR PUBLISHED WORKS.



RightsLink®

Home

Account
Info

Help



ACS Publications
Most Trusted. Most Cited. Most Read.

Title: A Computational Investigation of Small-Molecule Engagement of Hot Spots at Protein-Protein Interaction Interfaces

Author: David Xu, Yubing Si, Samy O. Meroueh

Publication: Journal of Chemical Information and Modeling

Publisher: American Chemical Society

Date: Sep 1, 2017

Copyright © 2017, American Chemical Society

Logged in as:

David Xu

LOGOUT

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK

CLOSE WINDOW

Copyright © 2019 Copyright Clearance Center, Inc. All Rights Reserved. [Privacy statement](#). [Terms and Conditions](#).
Comments? We would like to hear from you. E-mail us at customer@copyright.com



Title: Mimicking Intermolecular Interactions of Tight Protein-Protein Complexes for Small-Molecule Antagonists

Author: David Xu, Khuchtumur Bum-Erdene, Yubing Si, et al

Publication: ChemMedChem

Publisher: John Wiley and Sons

Date: Oct 23, 2017

© WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Logged in as:

David Xu

LOGOUT

Order Completed

Thank you for your order.

This Agreement between Mr. David Xu ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

[printable details](#)

License Number	4526120477059
License date	Feb 11, 2019
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	ChemMedChem
Licensed Content Title	Mimicking Intermolecular Interactions of Tight Protein-Protein Complexes for Small-Molecule Antagonists
Licensed Content Author	David Xu, Khuchtumur Bum-Erdene, Yubing Si, et al
Licensed Content Date	Oct 23, 2017
Licensed Content Volume	12
Licensed Content Issue	21
Licensed Content Pages	16
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Title of your thesis / dissertation	COMPUTATIONAL METHODS TO IDENTIFY AND TARGET DRUGGABLE BINDING SITES AT PROTEIN-PROTEIN INTERACTIONS IN THE HUMAN INTERACTOME
Expected completion date	Mar 2019
Expected size (number of pages)	200
Requestor Location	Mr. David Xu 410 W. 10th Street HS 5000 INDIANAPOLIS, IN 46202 United States Attn: Mr. David Xu
Publisher Tax ID	EU826007151



Title: Chemical Space Overlap with Critical Protein-Protein Interface Residues in Commercial and Specialized Small-Molecule Libraries

Author: Yubing Si, David Xu, Khuchtumur Bum-Erdene, et al

Publication: ChemMedChem

Publisher: John Wiley and Sons

Date: Dec 20, 2018

© WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Logged in as:

David Xu

LOGOUT

Order Completed

Thank you for your order.

This Agreement between Mr. David Xu ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

[printable details](#)

License Number	4526120732974
License date	Feb 11, 2019
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	ChemMedChem
Licensed Content Title	Chemical Space Overlap with Critical Protein-Protein Interface Residues in Commercial and Specialized Small-Molecule Libraries
Licensed Content Author	Yubing Si, David Xu, Khuchtumur Bum-Erdene, et al
Licensed Content Date	Dec 20, 2018
Licensed Content Volume	14
Licensed Content Issue	1
Licensed Content Pages	13
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Title of your thesis / dissertation	COMPUTATIONAL METHODS TO IDENTIFY AND TARGET DRUGGABLE BINDING SITES AT PROTEIN-PROTEIN INTERACTIONS IN THE HUMAN INTERACTOME
Expected completion date	Mar 2019
Expected size (number of pages)	200
Requestor Location	Mr. David Xu 410 W. 10th Street HS 5000 INDIANAPOLIS, IN 46202 United States Attn: Mr. David Xu
Publisher Tax ID	EU826007151

REFERENCES

1. Hanahan D, Weinberg RA. The Hallmarks of Cancer, *Cell* 2000;100:57-70.
2. Hammerman PS, Lawrence MS, Voet D et al. Comprehensive Genomic Characterization of Squamous Cell Lung Cancers, *Nature* 2012;489:519-525.
3. Weir BA, Woo MS, Getz G et al. Characterizing the Cancer Genome in Lung Adenocarcinoma, *Nature* 2007;450:893-898.
4. The Cancer Genome Atlas Research Network. Integrated Genomic Analyses of Ovarian Carcinoma, *Nature* 2011;474:609-615.
5. The Cancer Genome Atlas Network. Comprehensive Molecular Characterization of Human Colon and Rectal Cancer, *Nature* 2012;487:330-337.
6. The Cancer Genome Atlas Network. Comprehensive Molecular Portraits of Human Breast Tumours, *Nature* 2012;490:61-70.
7. The Cancer Genome Atlas Research Network. Comprehensive Molecular Characterization of Clear Cell Renal Cell Carcinoma, *Nature* 2013;499:43-49.
8. The Cancer Genome Atlas Research Network. Comprehensive Genomic Characterization of Squamous Cell Lung Cancers, *Nature* 2012;489:519-525.
9. The Cancer Genome Atlas Research Network. Comprehensive Molecular Profiling of Lung Adenocarcinoma, *Nature* 2014;511:543-550.
10. Bailey P, Chang DK, Nones K et al. Genomic Analyses Identify Molecular Subtypes of Pancreatic Cancer, *Nature* 2016;531:47-52.
11. Biankin AV, Waddell N, Kassahn KS et al. Pancreatic Cancer Genomes Reveal Aberrations in Axon Guidance Pathway Genes, *Nature* 2012;491:399-405.
12. Brennan CW, Verhaak RG, McKenna A et al. The Somatic Genomic Landscape of Glioblastoma, *Cell* 2013;155:462-477.
13. Parsons DW, Jones S, Zhang X et al. An Integrated Genomic Analysis of Human Glioblastoma Multiforme, *Science* 2008;321:1807-1812.
14. Kandoth C, McLellan MD, Vandin F et al. Mutational landscape and significance across 12 major cancer types, *Nature* 2013;502:333-339.
15. Vogelstein B, Papadopoulos N, Velculescu VE et al. Cancer Genome Landscapes, *Science* 2013;339:1546-1558.
16. Shi Z, Moulton J. Structural and Functional Impact of Cancer-Related Missense Somatic Mutations, *J. Mol. Biol.* 2011;413:495-512.
17. Ivanov AA, Khuri FR, Fu H. Targeting Protein-Protein Interactions as an Anticancer Strategy, *Trends Pharmacol. Sci.* 2013;34:393-400.

18. Ryan DP, Matthews JM. Protein-Protein Interactions in Human Disease, *Curr. Opin. Struct. Biol.* 2005;15:441-446.
19. Venkatesan K, Rual JF, Vazquez A et al. An Empirical Framework for Binary Interactome Mapping, *Nat. Methods* 2009;6:83-90.
20. Stelzl U, Worm U, Lalowski M et al. A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome, *Cell* 2005;122:957-968.
21. Rual JF, Venkatesan K, Hao T et al. Towards a Proteome-Scale Map of the Human Protein-Protein Interaction Network, *Nature* 2005;437:1173-1178.
22. Arkin MR, Tang Y, Wells JA. Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing Toward the Reality, *Chem. Biol.* 2014;21:1102-1114.
23. Wells JA, McClendon CL. Reaching for High-Hanging Fruit in Drug Discovery at Protein-Protein Interfaces, *Nature* 2007;450:1001-1009.
24. Hopkins AL, Groom CR. The Druggable Genome, *Nat. Rev. Drug Discovery* 2002;1:727-730.
25. Bahadur RP, Chakrabarti P, Rodier F et al. A Dissection of Specific and Non-Specific Protein-Protein Interfaces, *J. Mol. Biol.* 2004;336:943-955.
26. Perkins JR, Diboun I, Dessailly BH et al. Transient Protein-Protein Interactions: Structural, Functional, and Network Properties, *Structure* 2010;18:1233-1243.
27. 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.
28. Bogan AA, Thorn KS. Anatomy of Hot Spots in Protein Interfaces, *J. Mol. Biol.* 1998;280:1-9.
29. Cukuroglu E, Engin HB, Gursoy A et al. Hot spots in protein-protein interfaces: towards drug discovery, *Prog Biophys Mol Biol* 2014;116:165-173.
30. Weiss GA, Watanabe CK, Zhong A et al. Rapid Mapping of Protein Functional Epitopes by Combinatorial Alanine Scanning, *Proc. Natl. Acad. Sci. U. S. A.* 2000;97:8950-8954.
31. Brenke R, Kozakov D, Chuang GY et al. Fragment-Based Identification of Druggable 'Hot Spots' of Proteins Using Fourier Domain Correlation Techniques, *Bioinformatics* 2009;25:621-627.
32. Grosdidier S, Fernandez-Recio J. Identification of Hot-Spot Residues in Protein-Protein Interactions by Computational Docking, *BMC Bioinf.* 2008;9:447.
33. Clackson T, Wells JA. A Hot Spot of Binding Energy in a Hormone-Receptor Interface, *Science* 1995;267:383-386.

34. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces, *J Mol Biol* 1998;280:1-9.
35. Mohamed R, Degac J, Helms V. Composition of Overlapping Protein-Protein and Protein-Ligand Interfaces, *PLoS One* 2015;10:e0140965.
36. Sheinerman FB, Norel R, Honig B. Electrostatic Aspects of Protein-Protein Interactions, *Curr. Opin. Struct. Biol.* 2000;10:153-159.
37. Crowley PB, Golovin A. Cation-Pi Interactions in Protein-Protein Interfaces, *Proteins: Struct., Funct., Bioinf.* 2005;59:231-239.
38. Keskin O, Ma B, Nussinov R. Hot Regions in Protein-Protein Interactions: The Organization and Contribution of Structurally Conserved Hot Spot Residues, *J. Mol. Biol.* 2005;345:1281-1294.
39. Santos R, Ursu O, Gaulton A et al. A Comprehensive Map of Molecular Drug Targets, *Nat. Rev. Drug Discovery* 2017;16:19-34.
40. Gohlke H, Kiel C, Case DA. Insights into Protein-Protein Binding by Binding Free Energy Calculation and Free Energy Decomposition for the Ras-Raf and Ras-RalGDS Complexes, *J. Mol. Biol.* 2003;330:891-913.
41. Metz A, Pflieger C, Kopitz H et al. Hot Spots and Transient Pockets: Predicting the Determinants of Small-Molecule Binding to a Protein-Protein Interface, *J. Chem. Inf. Model.* 2012;52:120-133.
42. Scott DE, Bayly AR, Abell C et al. Small Molecules, Big Targets: Drug Discovery Faces the Protein-Protein Interaction Challenge, *Nat. Rev. Drug Discovery* 2016;15:533-550.
43. Pelay-Gimeno M, Glas A, Koch O et al. Structure-Based Design of Inhibitors of Protein-Protein Interactions: Mimicking Peptide Binding Epitopes, *Angew. Chem., Int. Ed. Engl.* 2015;54:8896-8927.
44. Azzarito V, Long K, Murphy NS et al. Inhibition of Alpha-Helix-Mediated Protein-Protein Interactions Using Designed Molecules, *Nat. Chem.* 2013;5:161-173.
45. Zhao Y, Aguilar A, Bernard D et al. Small-Molecule Inhibitors of the MDM2-p53 Protein-Protein Interaction (MDM2 Inhibitors) in Clinical Trials for Cancer Treatment, *J. Med. Chem.* 2015;58:1038-1052.
46. Hajduk PJ, Greer J. A Decade of Fragment-Based Drug Design: Strategic Advances and Lessons Learned, *Nat. Rev. Drug Discovery* 2007;6:211-219.
47. Petros AM, Dinges J, Augeri DJ et al. Discovery of a Potent Inhibitor of the Antiapoptotic Protein Bcl-xL from NMR and Parallel Synthesis, *J. Med. Chem.* 2006;49:656-663.

48. Wilson CG, Arkin MR. Small-Molecule Inhibitors of IL-2/IL-2R: Lessons Learned and Applied, *Curr. Top. Microbiol. Immunol.* 2011;348:25-59.
49. Davies TG, Wixted WE, Coyle JE et al. Monoacidic Inhibitors of the Kelch-like ECH-Associated Protein 1: Nuclear Factor Erythroid 2-Related Factor 2 (KEAP1:NRF2) Protein-Protein Interaction with High Cell Potency Identified by Fragment-Based Discovery, *J. Med. Chem.* 2016;59:3991-4006.
50. Geppert T, Bauer S, Hiss JA et al. Immunosuppressive Small Molecule Discovered by Structure-Based Virtual Screening for Inhibitors of Protein-Protein Interactions, *Angew. Chem., Int. Ed. Engl.* 2012;51:258-261.
51. Christ F, Voet A, Marchand A et al. Rational Design of Small-Molecule Inhibitors of the LEDGF/p75-Integrase Interaction and HIV Replication, *Nat. Chem. Biol.* 2010;6:442-448.
52. Zhou H, Chen J, Meagher JL et al. Design of Bcl-2 and Bcl-xL Inhibitors with Subnanomolar Binding Affinities Based upon a New Scaffold, *J. Med. Chem.* 2012;55:4664-4682.
53. Hopkins AL, Keseru GM, Leeson PD et al. The Role of Ligand Efficiency Metrics in Drug Discovery, *Nat. Rev. Drug Discovery* 2014;13:105-121.
54. Hopkins AL, Groom CR, Alex A. Ligand Efficiency: A Useful Metric for Lead Selection, *Drug Discovery Today* 2004;9:430-431.
55. Leeson PD, Springthorpe B. The Influence of Drug-Like Concepts on Decision-Making in Medicinal Chemistry, *Nat. Rev. Drug Discovery* 2007;6:881-890.
56. Schultes S, Graaf Cd, Haaksma EEJ et al. Ligand Efficiency as a Guide in Fragment Hit Selection and Optimization, *Drug Discovery Today: Technol.* 2010;7:e147-202.
57. Sattler M, Liang H, Nettlesheim D et al. Structure of Bcl-xL-Bak Peptide Complex: Recognition Between Regulators of Apoptosis, *Science* 1997;275:983-986.
58. Oltersdorf T, Elmore SW, Shoemaker AR et al. An Inhibitor of Bcl-2 Family Proteins Induces Regression of Solid Tumours, *Nature* 2005;435:677-681.
59. Lee EF, Czabotar PE, Smith BJ et al. Crystal Structure of ABT-737 Complexed with Bcl-xL: Implications for Selectivity of Antagonists of the Bcl-2 Family, *Cell Death Differ.* 2007;14:1711-1713.
60. Sleebs BE, Czabotar PE, Fairbrother WJ et al. Quinazoline Sulfonamides as Dual Binders of the Proteins B-Cell Lymphoma 2 and B-Cell Lymphoma Extra Long with Potent Proapoptotic Cell-Based Activity, *J. Med. Chem.* 2011;54:1914-1926.
61. Zhou H, Aguilar A, Chen J et al. Structure-Based Design of Potent Bcl-2/Bcl-xL Inhibitors with Strong in Vivo Antitumor Activity, *J. Med. Chem.* 2012;55:6149-6161.

62. Tao ZF, Hasvold L, Wang L et al. Discovery of a Potent and Selective BCL-XL Inhibitor with *in Vivo* Activity, *ACS Med. Chem. Lett.* 2014;5:1088-1093.
63. Pazgier M, Liu M, Zou G et al. Structural Basis for High-Affinity Peptide Inhibition of p53 Interactions with MDM2 and MDMX, *Proc. Natl. Acad. Sci. U. S. A.* 2009;106:4665-4670.
64. Vassilev LT, Vu BT, Graves B et al. In Vivo Activation of the p53 Pathway by Small-Molecule Antagonists of MDM2, *Science* 2004;303:844-848.
65. Grasberger BL, Lu T, Schubert C et al. Discovery and Cocrystal Structure of Benzodiazepinedione HDM2 Antagonists That Activate p53 in Cells, *J. Med. Chem.* 2005;48:909-912.
66. Raboisson P, Marugan JJ, Schubert C et al. Structure-Based Design, Synthesis, and Biological Evaluation of Novel 1,4-Diazepines as HDM2 Antagonists, *Bioorg. Med. Chem. Lett.* 2005;15:1857-1861.
67. Allen JG, Bourbeau MP, Wohlhieter GE et al. Discovery and Optimization of Chromenotriazolopyrimidines as Potent Inhibitors of the Mouse Double Minute 2-Tumor Protein 53 Protein-Protein Interaction, *J. Med. Chem.* 2009;52:7044-7053.
68. Popowicz GM, Czarna A, Wolf S et al. Structures of Low Molecular Weight Inhibitors Bound to MDMX and MDM2 Reveal New Approaches for p53-MDMX/MDM2 Antagonist Drug Discovery, *Cell Cycle* 2010;9:1104-1111.
69. Huang Y, Wolf S, Koes D et al. Exhaustive Fluorine Scanning Toward Potent p53-Mdm2 Antagonists, *ChemMedChem* 2012;7:49-52.
70. Miyazaki M, Naito H, Sugimoto Y et al. Synthesis and Evaluation of Novel Orally Active p53-MDM2 Interaction Inhibitors, *Bioorg. Med. Chem.* 2013;21:4319-4331.
71. Furet P, Chene P, De Pover A et al. The Central Valine Concept Provides an Entry in a New Class of Non Peptide Inhibitors of the p53-MDM2 Interaction, *Bioorg. Med. Chem. Lett.* 2012;22:3498-3502.
72. Rew Y, Sun D, Gonzalez-Lopez De Turiso F et al. Structure-Based Design of Novel Inhibitors of the MDM2-p53 Interaction, *J. Med. Chem.* 2012;55:4936-4954.
73. Anil B, Riedinger C, Endicott JA et al. The Structure of an MDM2-Nutlin-3a Complex Solved by the Use of a Validated MDM2 Surface-Entropy Reduction Mutant, *Acta Crystallogr., Sect. D: Biol. Crystallogr.* 2013;69:1358-1366.
74. Liu Z, Sun C, Olejniczak ET et al. Structural Basis for Binding of Smac/DIABLO to the XIAP BIR3 Domain, *Nature* 2000;408:1004-1008.
75. Sun H, Stuckey JA, Nikolovska-Coleska Z et al. Structure-Based Design, Synthesis, Evaluation, and Crystallographic Studies of Conformationally Constrained Smac Mimetics as

Inhibitors of the X-Linked Inhibitor of Apoptosis Protein (XIAP), *J. Med. Chem.* 2008;51:7169-7180.

76. Wist AD, Gu L, Riedl SJ et al. Structure-Activity Based Study of the Smac-Binding Pocket Within the BIR3 Domain of XIAP, *Bioorg. Med. Chem.* 2007;15:2935-2943.

77. Mastrangelo E, Cossu F, Milani M et al. Targeting the X-Linked Inhibitor of Apoptosis Protein through 4-Substituted Azabicyclo[5.3.0]alkane Smac Mimetics. Structure, Activity, and Recognition Principles, *J. Mol. Biol.* 2008;384:673-689.

78. Cossu F, Mastrangelo E, Milani M et al. Designing Smac-Mimetics as Antagonists of XIAP, cIAP1, and cIAP2, *Biochem. Biophys. Res. Commun.* 2009;378:162-167.

79. Ndubaku C, Varfolomeev E, Wang L et al. Antagonism of c-IAP and XIAP Proteins Is Required for Efficient Induction of Cell Death by Small-Molecule IAP Antagonists, *ACS Chem. Biol.* 2009;4:557-566.

80. Chessari G, Buck IM, Day JE et al. Fragment-Based Drug Discovery Targeting Inhibitor of Apoptosis Proteins: Discovery of a Non-Alanine Lead Series with Dual Activity Against cIAP1 and XIAP, *J. Med. Chem.* 2015;58:6574-6588.

81. Tilley JW, Chen L, Fry DC et al. Identification of a Small Molecule Inhibitor of the IL-2/IL-2R α Receptor Interaction Which Binds to IL-2, *J. Am. Chem. Soc.* 1997;119:7589-7590.

82. Arkin MR, Randal M, DeLano WL et al. Binding of Small Molecules to an Adaptive Protein-Protein Interface, *Proc. Natl. Acad. Sci. U. S. A.* 2003;100:1603-1608.

83. Thanos CD, Randal M, Wells JA. Potent Small-Molecule Binding to a Dynamic Hot Spot on IL-2, *J. Am. Chem. Soc.* 2003;125:15280-15281.

84. Thanos CD, DeLano WL, Wells JA. Hot-Spot Mimicry of a Cytokine Receptor by a Small Molecule, *Proc. Natl. Acad. Sci. U. S. A.* 2006;103:15422-15427.

85. Jung M, Philpott M, Muller S et al. Affinity Map of Bromodomain Protein 4 (BRD4) Interactions with the Histone H4 Tail and the Small Molecule Inhibitor JQ1, *J. Biol. Chem.* 2014;289:9304-9319.

86. Chung CW, Coste H, White JH et al. Discovery and Characterization of Small Molecule Inhibitors of the BET Family Bromodomains, *J. Med. Chem.* 2011;54:3827-3838.

87. Nicodeme E, Jeffrey KL, Schaefer U et al. Suppression of Inflammation by a Synthetic Histone Mimic, *Nature* 2010;468:1119-1123.

88. Sasaki K, Ito A, Yoshida M. Development of Live-Cell Imaging Probes for Monitoring Histone Modifications, *Bioorg. Med. Chem.* 2012;20:1887-1892.

89. Filippakopoulos P, Picaud S, Fedorov O et al. Benzodiazepines and Benzotriazepines as Protein Interaction Inhibitors Targeting Bromodomains of the BET Family, *Bioorg. Med. Chem.* 2012;20:1878-1886.
90. Dawson MA, Prinjha RK, Dittmann A et al. Inhibition of BET Recruitment to Chromatin as an Effective Treatment for MLL-Fusion Leukaemia, *Nature* 2011;478:529-533.
91. Zhang G, Liu R, Zhong Y et al. Down-Regulation of NF-KappaB Transcriptional Activity in HIV-Associated Kidney Disease by BRD4 Inhibition, *J. Biol. Chem.* 2012;287:28840-28851.
92. Picaud S, Wells C, Felletar I et al. RVX-208, an Inhibitor of BET Transcriptional Regulators with Selectivity for the Second Bromodomain, *Proc. Natl. Acad. Sci. U. S. A.* 2013;110:19754-19759.
93. Hugle M, Lucas X, Weitzel G et al. 4-Acyl Pyrrole Derivatives Yield Novel Vectors for Designing Inhibitors of the Acetyl-Lysine Recognition Site of BRD4(1), *J. Med. Chem.* 2016;59:1518-1530.
94. Sitkoff D, Sharp KA, Honig B. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models, *J. Phys. Chem.* 1994;98:1978-1988.
95. Adams JM, Cory S. The Bcl-2 Apoptotic Switch in Cancer Development and Therapy, *Oncogene* 2007;26:1324-1337.
96. Fesik SW. Promoting Apoptosis as a Strategy for Cancer Drug Discovery, *Nat. Rev. Cancer* 2005;5:876-885.
97. Petros AM, Olejniczak ET, Fesik SW. Structural Biology of the Bcl-2 Family of Proteins, *Biochim. Biophys. Acta* 2004;1644:83-94.
98. Oberstein A, Jeffrey PD, Shi Y. Crystal Structure of the Bcl-XL-Bcl-1 Peptide Complex: Bcl-1 Is a Novel BH3-Only Protein, *J. Biol. Chem.* 2007;282:13123-13132.
99. Kelekar A, Chang BS, Harlan JE et al. Bad Is a BH3 Domain-Containing Protein That Forms an Inactivating Dimer with Bcl-XL, *Mol. Cell. Biol.* 1997;17:7040-7046.
100. Ding J, Mooers BH, Zhang Z et al. After Embedding in Membranes Antiapoptotic Bcl-XL Protein Binds Both Bcl-2 Homology Region 3 and Helix 1 of Proapoptotic Bax Protein to Inhibit Apoptotic Mitochondrial Permeabilization, *J. Biol. Chem.* 2014;289:11873-11896.
101. Wade M, Li YC, Wahl GM. MDM2, MDMX and p53 in Oncogenesis and Cancer Therapy, *Nat. Rev. Cancer* 2013;13:83-96.
102. Moll UM, Petrenko O. The MDM2-p53 Interaction, *Mol. Cancer Res.* 2003;1:1001-1008.
103. Bottger A, Bottger V, Garcia-Echeverria C et al. Molecular Characterization of the hdm2-p53 Interaction, *J. Mol. Biol.* 1997;269:744-756.

104. Sun D, Li Z, Rew Y et al. Discovery of AMG 232, a Potent, Selective, and Orally Bioavailable MDM2-p53 Inhibitor in Clinical Development, *J. Med. Chem.* 2014;57:1454-1472.
105. Galban S, Duckett CS. XIAP as a Ubiquitin Ligase in Cellular Signaling, *Cell Death Differ.* 2010;17:54-60.
106. Abhari BA, Davoodi J. A Mechanistic Insight into SMAC Peptide Interference with XIAP-Bir2 Inhibition of Executioner Caspases, *J. Mol. Biol.* 2008;381:645-654.
107. Srinivasula SM, Hegde R, Saleh A et al. A Conserved XIAP-Interaction Motif in Caspase-9 and Smac/DIABLO Regulates Caspase Activity and Apoptosis, *Nature* 2001;410:112-116.
108. Wu G, Chai J, Suber TL et al. Structural Basis of IAP Recognition by Smac/DIABLO, *Nature* 2000;408:1008-1012.
109. Liao W, Lin JX, Leonard WJ. Interleukin-2 at the Crossroads of Effector Responses, Tolerance, and Immunotherapy, *Immunity* 2013;38:13-25.
110. Mott HR, Baines BS, Hall RM et al. The Solution Structure of the F42A Mutant of Human Interleukin 2, *J. Mol. Biol.* 1995;247:979-994.
111. Wang Z, Zheng Z, Sun L et al. Substitutions at the Glu62 Residue of Human Interleukin-2 Differentially Affect Its Binding to the Alpha Chain and the Beta Gamma Complex of the Interleukin-2 Receptor, *Eur. J. Immunol.* 1995;25:1212-1216.
112. Sauve K, Nachman M, Spence C et al. Localization in Human Interleukin 2 of the Binding Site to the Alpha Chain (p55) of the Interleukin 2 Receptor, *Proc. Natl. Acad. Sci. U. S. A.* 1991;88:4636-4640.
113. Filippakopoulos P, Picaud S, Mangos M et al. Histone Recognition and Large-Scale Structural Analysis of the Human Bromodomain Family, *Cell* 2012;149:214-231.
114. Filippakopoulos P, Knapp S. Targeting Bromodomains: Epigenetic Readers of Lysine Acetylation, *Nat. Rev. Drug Discovery* 2014;13:337-356.
115. Dixon SL, Smondryev AM, Knoll EH et al. PHASE: A New Engine for Pharmacophore Perception, 3D QSAR Model Development, and 3D Database Screening: 1. Methodology and Preliminary Results, *J. Comput.-Aided Mol. Des.* 2006;20:647-671.
116. Dixon SL, Smondryev AM, Rao SN. PHASE: A Novel Approach to Pharmacophore Modeling and 3D Database Searching, *Chem. Biol. Drug Des.* 2006;67:370-372.
117. Hunenberger PH, Mark AE, van Gunsteren WF. Fluctuation and Cross-Correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations, *J. Mol. Biol.* 1995;252:492-503.
118. Basse MJ, Betzi S, Bourgeas R et al. 2P2Idb: A Structural Database Dedicated to Orthosteric Modulation of Protein-Protein Interactions, *Nucleic Acids Res.* 2013;41:D824-827.

119. UniProt Consortium. Uniprot: A Hub for Protein Information, *Nucleic Acids Res.* 2015;43:D204-212.
120. Wang R, Fang X, Lu Y et al. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures, *J. Med. Chem.* 2004;47:2977-2980.
121. Hu L, Benson ML, Smith RD et al. Binding MOAD (Mother Of All Databases), *Proteins: Struct., Funct., Bioinf.* 2005;60:333-340.
122. Liu T, Lin Y, Wen X et al. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities, *Nucleic Acids Res.* 2007;35:D198-201.
123. Jacobson MP, Pincus DL, Rapp CS et al. A Hierarchical Approach to All-Atom Protein Loop Prediction, *Proteins: Struct., Funct., Bioinf.* 2004;55:351-367.
124. Olsson MHM, Søndergaard CR, Rostkowski M et al. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions, *J. Chem. Theory Comput.* 2011;7:525-537.
125. Shelley JC, Cholleti A, Frye LL et al. Epik: A Software Program for pKa Prediction and Protonation State Generation for Drug-Like Molecules, *J. Comput.-Aided Mol. Des.* 2007;21:681-691.
126. Case DA, Berryman JT, Betz RM et al. AMBER 2016. San Francisco: University of California, 2016.
127. Jakalian A, Jack DB, Bayly CI. Fast, Efficient Generation of High-quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation, *J. Comput. Chem.* 2002;23:1623-1641.
128. Wang J, Wolf RM, Caldwell JW et al. Development and Testing of a General Amber Force Field, *J. Comput. Chem.* 2004;25:1157-1174.
129. Wang J, Wang W, Kollman PA et al. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations, *J. Mol. Graphics Modell.* 2006;25:247-260.
130. Jorgensen WL, Chandrasekhar J, Madura JD et al. Comparison of Simple Potential Functions for Simulating Liquid Water, *J. Chem. Phys.* 1983;79:926-935.
131. Maier JA, Martinez C, Kasavajhala K et al. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB, *J. Chem. Theory Comput.* 2015;11:3696-3713.
132. Ryckaert JP, Ciccotti G, Berendsen JJC. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes, *J. Comput. Phys.* 1977;23:327-341.
133. Darden T, York D, Pedersen L. Particle Mesh Ewald: An $N \cdot \log(N)$ Method for Ewald Sums in Large Systems, *J. Chem. Phys.* 1993;98:10089-10092.

134. Roe DR, Cheatham TE, 3rd. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data, *J. Chem. Theory Comput.* 2013;9:3084-3095.
135. Still WC, Tempczyk A, Hawley RC et al. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics, *J. Am. Chem. Soc.* 1990;112:6127-6129.
136. Miller BR, 3rd, McGee TD, Jr., Swails JM et al. MMPBSA.py: An Efficient Program for End-State Free Energy Calculations, *J. Chem. Theory Comput.* 2012;8:3314-3321.
137. Onufriev A, Bashford D, Case DA. Exploring Protein Native States and Large-scale Conformational Changes with a Modified Generalized Born Model, *Proteins: Struct., Funct., Bioinf.* 2004;55:383-394.
138. Feig M, Onufriev A, Lee MS et al. Performance Comparison of Generalized Born and Poisson Methods in the Calculation of Electrostatic Solvation Energies for Protein Structures, *J. Comput. Chem.* 2004;25:265-284.
139. Brooks B, Karplus M. Harmonic Dynamics of Proteins: Normal Modes and Fluctuations in Bovine Pancreatic Trypsin Inhibitor, *Proc. Natl. Acad. Sci. U. S. A.* 1983;80:6571-6575.
140. Kollman PA, Massova I, Reyes C et al. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models, *Acc. Chem. Res.* 2000;33:889-897.
141. van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation, *IEEE Comput. Sci. Eng.* 2011;13:22-30.
142. Kastiris PL, Moal IH, Hwang H et al. A Structure-Based Benchmark for Protein-Protein Binding Affinity, *Protein Sci.* 2011;20:482-491.
143. Tovar C, Graves B, Packman K et al. MDM2 Small-Molecule Antagonist RG7112 Activates p53 Signaling and Regresses Human Tumors in Preclinical Cancer Models, *Cancer Res.* 2013;73:2587-2597.
144. Ding Q, Zhang Z, Liu JJ et al. Discovery of RG7388, a Potent and Selective p53-MDM2 Inhibitor in Clinical Development, *J. Med. Chem.* 2013;56:5979-5983.
145. Roehrl MH, Kang S, Aramburu J et al. Selective Inhibition of Calcineurin-NFAT Signaling by Blocking Protein-Protein Interaction with Small Organic Molecules, *Proc. Natl. Acad. Sci. U. S. A.* 2004;101:7554-7559.
146. Lepourcelet M, Chen YN, France DS et al. Small-Molecule Antagonists of the Oncogenic Tcf/Beta-Catenin Protein Complex, *Cancer Cell* 2004;5:91-102.
147. Mukherjee P, Desai P, Zhou YD et al. Targeting the BH3 Domain Mediated Protein-Protein Interaction of Bcl-xL Through Virtual Screening, *J. Chem. Inf. Model.* 2010;50:906-923.

148. Hoggard LR, Zhang Y, Zhang M et al. Rational Design of Selective Small-Molecule Inhibitors for Beta-Catenin/B-Cell Lymphoma 9 Protein-Protein Interactions, *J. Am. Chem. Soc.* 2015;137:12249-12260.
149. Rognan D. Rational Design of Protein-Protein Interaction Inhibitors, *MedChemComm* 2015;6:51-60.
150. Lu Y, Nikolovska-Coleska Z, Fang X et al. Discovery of a Nanomolar Inhibitor of the Human Murine Double Minute 2 (MDM2)-p53 Interaction Through an Integrated, Virtual Database Screening Strategy, *J. Med. Chem.* 2006;49:3759-3762.
151. Tortorella P, Laghezza A, Durante M et al. An Effective Virtual Screening Protocol To Identify Promising p53-MDM2 Inhibitors, *J. Chem. Inf. Model.* 2016;56:1216-1227.
152. Wang L, Li L, Zhou ZH et al. Structure-Based Virtual Screening and Optimization of Modulators Targeting Hsp90-Cdc37 Interaction, *Eur. J. Med. Chem.* 2017;136:63-73.
153. Melagraki G, Ntougkos E, Rinotas V et al. Cheminformatics-Aided Discovery of Small-Molecule Protein-Protein Interaction (PPI) Dual Inhibitors of Tumor Necrosis Factor (TNF) and Receptor Activator of NF-KappaB Ligand (RANKL), *PLoS Comput. Biol.* 2017;13:e1005372.
154. Gessier F, Kallen J, Jacoby E et al. Discovery of Dihydroisoquinolinone Derivatives as Novel Inhibitors of the p53-MDM2 Interaction with a Distinct Binding Mode, *Bioorg. Med. Chem. Lett.* 2015;25:3621-3625.
155. Leung KH, Liu LJ, Lin S et al. Discovery of a Small-Molecule Inhibitor of STAT3 by Ligand-Based Pharmacophore Screening, *Methods* 2015;71:38-43.
156. Li H, Xiao H, Lin L et al. Drug Design Targeting Protein-Protein Interactions (PPIs) Using Multiple Ligand Simultaneous Docking (MLSD) and Drug Repositioning: Discovery of Raloxifene and Bazedoxifene as Novel Inhibitors of IL-6/GP130 Interface, *J. Med. Chem.* 2014;57:632-641.
157. Magdolen V, Rettenberger P, Koppitz M et al. Systematic Mutational Analysis of the Receptor-Binding Region of the Human Urokinase-Type Plasminogen Activator, *Eur. J. Biochem.* 1996;237:743-751.
158. Gardsvoll H, Ploug M. Mapping of the Vitronectin-Binding Site on the Urokinase Receptor: Involvement of a Coherent Receptor Interface Consisting of Residues from Both Domain I and the Flanking Interdomain Linker Region, *J. Biol. Chem.* 2007;282:13561-13572.
159. Madsen CD, Ferraris GM, Andolfo A et al. uPAR-Induced Cell Adhesion and Migration: Vitronectin Provides the Key, *J. Cell Biol.* 2007;177:927-939.
160. Huai Q, Zhou A, Lin L et al. Crystal Structures of Two Human Vitronectin, Urokinase and Urokinase Receptor Complexes, *Nat. Struct. Mol. Biol.* 2008;15:422-423.

161. Gardsvoll H, Dano K, Ploug M. Mapping Part of the Functional Epitope for Ligand Binding on the Receptor for Urokinase-Type Plasminogen Activator by Site-Directed Mutagenesis, *J. Biol. Chem.* 1999;274:37995-38003.
162. Huai Q, Mazar AP, Kuo A et al. Structure of Human Urokinase Plasminogen Activator in Complex with Its Receptor, *Science* 2006;311:656-659.
163. Llinas P, Le Du MH, Gardsvoll H et al. Crystal Structure of the Human Urokinase Plasminogen Activator Receptor Bound to an Antagonist Peptide, *EMBO J.* 2005;24:1655-1663.
164. Gårdsvoll H, Gilquin B, Le Du MH et al. Characterization of the Functional Epitope on the Urokinase Receptor. Complete Alanine Scanning Mutagenesis Supplemented by Chemical Cross-Linking, *J. Biol. Chem.* 2006;281:19260-19272.
165. Khanna M, Wang F, Jo I et al. Targeting Multiple Conformations Leads to Small Molecule Inhibitors of the uPAR·uPA Protein-Protein Interaction that Block Cancer Cell Invasion, *ACS Chem. Biol.* 2011;6:1232-1243.
166. McCallum MM, Nandhikonda P, Temmer JJ et al. High-Throughput Identification of Promiscuous Inhibitors from Screening Libraries with the Use of a Thiol-Containing Fluorescent Probe, *J. Biomol. Screening* 2013;18:705-713.
167. Liu D, Zhou D, Wang B et al. A New Class of Orthosteric uPAR·uPA Small-Molecule Antagonists Are Allosteric Inhibitors of the uPAR·Vitronectin Interaction, *ACS Chem. Biol.* 2015;10:1521-1534.
168. Liu D, Xu D, Liu M et al. Small Molecules Engage Hot Spots Through Cooperative Binding to Inhibit a Tight Protein-Protein Interaction, *Biochemistry* 2017;56:1768-1784.
169. Irwin JJ, Sterling T, Mysinger MM et al. ZINC: A Free Tool to Discover Chemistry for Biology, *J. Chem. Inf. Model.* 2012;52:1757-1768.
170. Baell JB, Holloway GA. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays, *J. Med. Chem.* 2010;53:2719-2740.
171. Walters WP, Murcko AA, Murcko MA. Recognizing Molecules with Drug-Like Properties, *Curr. Opin. Chem. Biol.* 1999;3:384-387.
172. 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;31:455-461.
173. Sastry GM, Adzhigirey M, Day T et al. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments, *J. Comput.-Aided Mol. Des.* 2013;27:221-234.

174. Bondi A. van der Waals Volumes and Radii, *J. Phys. Chem.* 1964;68:441-451.
175. Jacobsen B, Gardsvoll H, Juhl Funch G et al. One-Step Affinity Purification of Recombinant Urokinase-Type Plasminogen Activator Receptor Using a Synthetic Peptide Developed by Combinatorial Chemistry, *Protein Expression Purif.* 2007;52:286-296.
176. Soares KM, Blackmon N, Shun TY et al. Profiling the NIH Small Molecule Repository for Compounds That Generate H₂O₂ by Redox Cycling in Reducing Environments, *Assay Drug Dev. Technol.* 2010;8:152-174.
177. Johnston PA, Soares KM, Shinde SN et al. Development of a 384-Well Colorimetric Assay to Quantify Hydrogen Peroxide Generated by the Redox Cycling of Compounds in the Presence of Reducing Agents, *Assay Drug Dev. Technol.* 2008;6:505-518.
178. Xu D, Bum-Erdene K, Si Y et al. Mimicking Intermolecular Interactions of Tight Protein-Protein Complexes for Small-Molecule Antagonists, *ChemMedChem* 2017;12:1794-1809.
179. Rullo AF, Fitzgerald KJ, Muthusamy V et al. Re-engineering the Immune Response to Metastatic Cancer: Antibody-Recruiting Small Molecules Targeting the Urokinase Receptor, *Angew. Chem., Int. Ed. Engl.* 2016;55:3642-3646.
180. Labbe CM, Kuenemann MA, Zarzycka B et al. iPPI-DB: An Online Database of Modulators of Protein-Protein Interactions, *Nucleic Acids Res.* 2016;44:D542-547.
181. Bourgeois R, Basse MJ, Morelli X et al. Atomic Analysis of Protein-Protein Interfaces With Known Inhibitors: The 2P2I Database, *PLoS One* 2010;5:e9598.
182. Higuero AP, Jubb H, Blundell TL. TIMBAL v2: Update of a Database Holding Small Molecules Modulating Protein-Protein Interactions, *Database* 2013;2013:bat039.
183. Villoutreix BO, Kuenemann MA, Poyet JL et al. Drug-Like Protein-Protein Interaction Modulators: Challenges and Opportunities for Drug Discovery and Chemical Biology, *Mol. Inf.* 2014;33:414-437.
184. Jin X, Lee K, Kim NH et al. Natural Products Used as a Chemical Library for Protein-Protein Interaction Targeted Drug Discovery, *J. Mol. Graphics Modell.* 2018;79:46-58.
185. Reynes C, Host H, Camproux AC et al. Designing Focused Chemical Libraries Enriched in Protein-Protein Interaction Inhibitors Using Machine-Learning Methods, *PLoS Comput. Biol.* 2010;6:e1000695.
186. Xu D, Si Y, Meroueh SO. A Computational Investigation of Small-Molecule Engagement of Hot Spots at Protein-Protein Interaction Interfaces, *J. Chem. Inf. Model.* 2017;57:2250-2272.
187. Spring DR. Diversity-Oriented Synthesis; A Challenge for Synthetic Chemists, *Org. Biomol. Chem.* 2003;1:3867-3870.

188. Spandl RJ, Bender A, Spring DR. Diversity-Oriented Synthesis; A Spectrum of Approaches and Results, *Org. Biomol. Chem.* 2008;6:1149-1158.
189. Cordier C, Morton D, Murrison S et al. Natural Products as an Inspiration in the Diversity-Oriented Synthesis of Bioactive Compound Libraries, *Nat. Prod. Rep.* 2008;25:719-737.
190. Tan DS. Diversity-Oriented Synthesis: Exploring the Intersections Between Chemistry and Biology, *Nat. Chem. Biol.* 2005;1:74-84.
191. Lenci E, Menchi G, Trabocchi A. Carbohydrates in Diversity-Oriented Synthesis: Challenges and Opportunities, *Org. Biomol. Chem.* 2016;14:808-825.
192. CJ OC, Beckmann HS, Spring DR. Diversity-Oriented Synthesis: Producing Chemical Tools for Dissecting Biology, *Chem. Soc. Rev.* 2012;41:4444-4456.
193. Comer E, Duvall JR, duPont Lee Mt. Utilizing Diversity-Oriented Synthesis in Antimicrobial Drug Discovery, *Future Med. Chem.* 2014;6:1927-1942.
194. Gerry CJ, Schreiber SL. Chemical Probes and Drug Leads From Advances in Synthetic Planning and Methodology, *Nat. Rev. Drug Discovery* 2018;17:333-352.
195. Dandapani S, Marcaurelle LA. Grand Challenge Commentary: Accessing New Chemical Space for 'Undruggable' Targets, *Nat. Chem. Biol.* 2010;6:861-863.
196. Galloway WR, Bender A, Welch M et al. The Discovery of Antibacterial Agents Using Diversity-Oriented Synthesis, *Chem. Commun. (Cambridge, U. K.)* 2009;0:2446-2462.
197. Galloway WR, Spring DR. Is Synthesis the Main Hurdle for the Generation of Diversity in Compound Libraries for Screening?, *Expert Opin. Drug Discovery* 2009;4:467-472.
198. Haggarty SJ. The Principle of Complementarity: Chemical Versus Biological Space, *Curr. Opin. Chem. Biol.* 2005;9:296-303.
199. Chevillard F, Kolb P. SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability, *J. Chem. Inf. Model.* 2015;55:1824-1835.
200. Marcaurelle LA, Comer E, Dandapani S et al. An Aldol-Based Build/Couple/Pair Strategy for the Synthesis of Medium- and Large-Sized Rings: Discovery of Macrocyclic Histone Deacetylase Inhibitors, *J. Am. Chem. Soc.* 2010;132:16962-16976.
201. Leeson PD. Molecular Inflation, Attrition and the Rule of Five, *Adv. Drug Delivery Rev.* 2016;101:22-33.
202. Veber DF, Johnson SR, Cheng HY et al. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates, *J. Med. Chem.* 2002;45:2615-2623.
203. Leeson PD, Davis AM. Time-Related Differences in the Physical Property Profiles of Oral Drugs, *J. Med. Chem.* 2004;47:6338-6348.

204. Baell J, Congreve M, Leeson P et al. Ask the Experts: Past, Present and Future of the Rule of Five, *Future Med. Chem.* 2013;5:745-752.
205. Chen L, Chan SW, Zhang X et al. Structural Basis of YAP Recognition by TEAD4 in the Hippo Pathway, *Genes Dev.* 2010;24:290-300.
206. Li Z, Zhao B, Wang P et al. Structural Insights Into the YAP and TEAD Complex, *Genes Dev.* 2010;24:235-240.
207. Van Petegem F, Duderstadt KE, Clark KA et al. Alanine-Scanning Mutagenesis Defines a Conserved Energetic Hotspot in the CaV α 1 AID-CaV β Interaction Site That Is Critical for Channel Modulation, *Structure* 2008;16:280-294.
208. Sterling T, Irwin JJ. ZINC 15--Ligand Discovery for Everyone, *J. Chem. Inf. Model.* 2015;55:2324-2337.
209. Duan JX, Dixon SL, Lowrie JF et al. Analysis and Comparison of 2D Fingerprints: Insights into Database Screening Performance Using Eight Fingerprint Methods, *J. Mol. Graphics Modell.* 2010;29:157-170.
210. Sastry M, Lowrie JF, Dixon SL et al. Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments, *J. Chem. Inf. Model.* 2010;50:771-784.
211. Irwin JJ, Shoichet BK. ZINC-- a Free Database of Commercially Available Compounds for Virtual Screening, *J. Chem. Inf. Model.* 2005;45:177-182.
212. Banks JL, Beard HS, Cao Y et al. Integrated Modeling Program, Applied Chemical Theory (IMPACT), *J. Comput. Chem.* 2005;26:1752-1780.
213. Greenwood JR, Calkins D, Sullivan AP et al. Towards the Comprehensive, Rapid, and Accurate Prediction of the Favorable Tautomeric States of Drug-Like Molecules in Aqueous Solution, *J. Comput.-Aided Mol. Des.* 2010;24:591-604.
214. Friesner RA, Banks JL, Murphy RB et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy, *J. Med. Chem.* 2004;47:1739-1749.
215. Halgren TA, Murphy RB, Friesner RA et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening, *J. Med. Chem.* 2004;47:1750-1759.
216. Weinstein JN, Collisson EA, Mills GB et al. The Cancer Genome Atlas Pan-Cancer Analysis Project, *Nat. Genet.* 2013;45:1113-1120.
217. The Cancer Genome Atlas Research Network. Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways, *Nature* 2008;455:1061-1068.

218. Zhao S, Fung-Leung WP, Bittner A et al. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells, *PLoS One* 2014;9:e78644.
219. Spruance SL, Reid JE, Grace M et al. Hazard Ratio in Clinical Trials, *Antimicrob. Agents Chemother.* 2004;48:2787-2792.
220. Zhang J, Yang PL, Gray NS. Targeting Cancer with Small Molecule Kinase Inhibitors, *Nat. Rev. Cancer* 2009;9:28-39.
221. Halgren TA. Identifying and Characterizing Binding Sites and Assessing Druggability, *J. Chem. Inf. Model.* 2009;49:377-389.
222. Papadakis AI, Sun C, Knijnenburg TA et al. SMARCE1 Suppresses EGFR Expression and Controls Responses to MET and ALK Inhibitors in Lung Cancer, *Cell Res.* 2015;25:445-458.
223. Nussinov R, Tsai C-J. Allosteric in Disease and in Drug Discovery, *Cell* 2013;153:293-305.
224. DeLaBarre B, Gross S, Fang C et al. Full-Length Human Glutaminase in Complex with an Allosteric Inhibitor, *Biochemistry* 2011;50:10764-10770.
225. Wu WI, Voegtli WC, Sturgis HL et al. Crystal Structure of Human AKT1 with an Allosteric Inhibitor Reveals a New Mode of Kinase Inhibition, *PLoS One* 2010;5:e12913.
226. Li L, Uversky VN, Dunker AK et al. A Computational Investigation of Allosteric in the Catabolite Activator Protein, *J. Am. Chem. Soc.* 2007;129:15668-15676.
227. UniProt Consortium. Reorganizing the Protein Space at the Universal Protein Resource (UniProt), *Nucleic Acids Res.* 2012;40:D71-75.
228. Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: A Resource of Catalytic Sites and Residues Identified in Enzymes Using Structural Data, *Nucleic Acids Res.* 2004;32:D129-133.
229. Roberts SA, Lawrence MS, Klimczak LJ et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers, *Nat. Genet.* 2013;45:970-976.
230. Zarrabi K, Dufour A, Li J et al. Inhibition of Matrix Metalloproteinase 14 (MMP-14)-Mediated Cancer Cell Migration, *J. Biol. Chem.* 2011;286:33167-33177.
231. Valencia K, Ormazabal C, Zanduetta C et al. Inhibition of Collagen Receptor Discoidin Domain Receptor-1 (DDR1) Reduces Cell Survival, Homing, and Colonization in Lung Cancer Bone Metastasis, *Clin. Cancer Res.* 2012;18:969-980.
232. Garten A, Petzold S, Korner A et al. Nampt: Linking NAD Biology, Metabolism and Cancer, *Trends Endocrinol. Metab.* 2009;20:130-138.
233. Goodey NM, Benkovic SJ. Allosteric Regulation and Catalysis Emerge via a Common Route, *Nat. Chem. Biol.* 2008;4:474-482.

234. Choi Y, Seeliger MA, Panjarian SB et al. N-Myristoylated c-Abl Tyrosine Kinase Localizes to the Endoplasmic Reticulum Upon Binding to an Allosteric Inhibitor, *J. Biol. Chem.* 2009;284:29005-29014.
235. Ostrem JM, Peters U, Sos ML et al. K-Ras(G12C) Inhibitors Allosterically Control GTP Affinity and Effector Interactions, *Nature* 2013;503:548-551.
236. Zhang QC, Petrey D, Garzon JI et al. PrePPI: A Structure-Informed Database of Protein-Protein Interactions, *Nucleic Acids Res.* 2013;41:D828-833.
237. Li F, Zhang Y, Wu C. Integrin-Linked Kinase Is Localized to Cell-Matrix Focal Adhesions but Not Cell-Cell Adhesion Sites and the Focal Adhesion Localization of Integrin-Linked Kinase Is Regulated by the PINCH-Binding ANK Repeats, *J. Cell Sci.* 1999;112 (Pt 24):4589-4599.
238. Ahmed S, Lee J, Kozma R et al. A Novel Functional Target for Tumor-Promoting Phorbol Esters and Lysophosphatidic Acid. The p21rac-GTPase Activating Protein N-Chimaerin, *J. Biol. Chem.* 1993;268:10709-10712.
239. Kanehisa M, Goto S, Sato Y et al. KEGG for Integration and Interpretation of Large-Scale Molecular Data Sets, *Nucleic Acids Res.* 2012;40:D109-114.
240. Yildirim MA, Goh KI, Cusick ME et al. Drug-Target Network, *Nat. Biotechnol.* 2007;25:1119-1126.
241. Goh KI, Cusick ME, Valle D et al. The Human Disease Network, *Proc. Natl. Acad. Sci. U. S. A.* 2007;104:8685-8690.
242. Campillos M, Kuhn M, Gavin AC et al. Drug Target Identification Using Side-Effect Similarity, *Science* 2008;321:263-266.
243. Lamb J, Crawford ED, Peck D et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease, *Science* 2006;313:1929-1935.
244. Taylor IW, Linding R, Warde-Farley D et al. Dynamic Modularity in Protein Interaction Networks Predicts Breast Cancer Outcome, *Nat. Biotechnol.* 2009;27:199-204.
245. Barabasi AL, Gulbahce N, Loscalzo J. Network Medicine: A Network-Based Approach to Human Disease, *Nat. Rev. Genet.* 2011;12:56-68.
246. Chen T, Sun Y, Ji P et al. Topoisomerase IIalpha in Chromosome Instability and Personalized Cancer Therapy, *Oncogene* 2014;34:4019-4031.
247. Kandath C, McLellan MD, Vandin F et al. Mutational Landscape and Significance Across 12 Major Cancer Types, *Nature* 2013;502:333-339.
248. Karakas B, Bachman KE, Park BH. Mutation of the PIK3CA Oncogene in Human Cancers, *Br. J. Cancer* 2006;94:455-459.

249. Rodriguez D, Ramsay AJ, Quesada V et al. Functional Analysis of Sucrase-Isomaltase Mutations From Chronic Lymphocytic Leukemia Patients, *Hum. Mol. Genet.* 2013;22:2273-2282.
250. Rodriguez-Escudero I, Oliver MD, Andres-Pons A et al. A Comprehensive Functional Analysis of PTEN Mutations: Implications in Tumor- and Autism-Related Syndromes, *Hum. Mol. Genet.* 2011;20:4132-4142.
251. Sun C, Wang L, Huang S et al. Reversible and Adaptive Resistance to BRAF(V600E) Inhibition in Melanoma, *Nature* 2014;508:118-122.
252. Ascierto PA, Kirkwood JM, Grob JJ et al. The Role of BRAF V600 Mutation in Melanoma, *J. Transl. Med.* 2012;10:85.
253. Cantwell-Dorris ER, O'Leary JJ, Sheils OM. BRAFV600E: Implications for Carcinogenesis and Molecular Therapy, *Mol. Cancer Ther.* 2011;10:385-394.
254. Guo Y, Zhang X, Yang M et al. Functional Evaluation of Missense Variations in the Human MAD1L1 and MAD2L1 Genes and Their Impact on Susceptibility to Lung Cancer, *J. Med. Genet.* 2010;47:616-622.
255. Tomimatsu N, Mukherjee B, Catherine Hardebeck M et al. Phosphorylation of EXO1 by CDKs 1 and 2 Regulates DNA End Resection and Repair Pathway Choice, *Nat. Commun.* 2014;5:3561.
256. Schmidtke P, Souaille C, Estienne F et al. Large-Scale Comparison of Four Binding Site Detection Algorithms, *J. Chem. Inf. Model.* 2010;50:2191-2200.
257. Halgren T. New Method for Fast and Accurate Binding-site Identification and Analysis, *Chem. Biol. Drug Des.* 2007;69:146-148.
258. Jubb H, Blundell TL, Ascher DB. Flexibility and Small Pockets at Protein-Protein Interfaces: New Insights Into Druggability, *Prog. Biophys. Mol. Biol.* 2015;119:2-9.
259. Nussinov R, Tsai C-J. The Different Ways Through Which Specificity Works in Orthosteric and Allosteric Drugs, *Curr. Pharm. Des.* 2012;18:1311-1316.
260. Heise CE, Murray J, Augustyn KE et al. Mechanistic and Structural Understanding of Uncompetitive Inhibitors of Caspase-6, *PloS One* 2012;7:e50864.
261. Orlicky S, Tang XJ, Neduva V et al. An Allosteric Inhibitor of Substrate Recognition by the SCFCdc4 Ubiquitin Ligase, *Nat. Biotechnol.* 2010;28:733-737.
262. Neklesa TK, Tae HS, Schneekloth AR et al. Small-Molecule Hydrophobic Tagging-Induced Degradation of HaloTag Fusion Proteins, *Nat. Chem. Biol.* 2011;7:538-543.
263. Cheng F, Jia P, Wang Q et al. Studying Tumorigenesis Through Network Evolution and Somatic Mutational Perturbations in the Cancer Interactome, *Mol. Biol. Evol.* 2014;31:2156-2169.

264. Zhao J, Cheng F, Wang Y et al. Systematic Prioritization of Druggable Mutations in Approximately 5000 Genomes Across 16 Cancer Types Using a Structural Genomics-Based Approach, *Mol. Cell. Proteomics* 2016;15:642-656.
265. Vuong H, Cheng F, Lin CC et al. Functional Consequences of Somatic Mutations in Cancer Using Protein Pocket-Based Prioritization Approach, *Genome Med.* 2014;6:81.
266. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data, *Bioinformatics* 2010;26:139-140.
267. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2013.
268. Rose PW, Beran B, Bi C et al. The RCSB Protein Data Bank: Redesigned Web Site and Web Services, *Nucleic Acids Res.* 2011;39:D392-401.
269. Altschul SF, Gish W, Miller W et al. Basic Local Alignment Search Tool, *J. Mol. Biol.* 1990;215:403-410.
270. Fu L, Niu B, Zhu Z et al. CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data, *Bioinformatics* 2012;28:3150-3152.
271. Jacobson MP, Friesner RA, Xiang Z et al. On the Role of the Crystal Environment in Determining Protein Side-Chain Conformations, *J Mol. Biol.* 2002;320:597-608.
272. Kabsch W, Sander C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features, *Biopolymers* 1983;22:2577-2637.
273. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function Using NetworkX. In: *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Pasadena, CA, 2008, p. 11-15.
274. Bader GD, Donaldson I, Wolting C et al. BIND--The Biomolecular Interaction Network Database, *Nucleic Acids Res.* 2001;29:242-245.
275. Stark C, Breitkreutz BJ, Reguly T et al. BioGRID: A General Repository for Interaction Datasets, *Nucleic Acids Res.* 2006;34:D535-539.
276. Xenarios I, Salwinski L, Duan XJ et al. DIP, the Database of Interacting Proteins: A Research Tool for Studying Cellular Networks of Protein Interactions, *Nucleic Acids Res.* 2002;30:303-305.
277. Peri S, Navarro JD, Kristiansen TZ et al. Human Protein Reference Database as a Discovery Resource for Proteomics, *Nucleic Acids Res.* 2004;32:D497-501.
278. Hermjakob H, Montecchi-Palazzi L, Lewington C et al. IntAct: An Open Source Molecular Interaction Database, *Nucleic Acids Res.* 2004;32:D452-455.

279. Chatr-aryamontri A, Ceol A, Palazzi LM et al. MINT: The Molecular INTeraction Database, *Nucleic Acids Res.* 2007;35:D572-574.
280. Croft D, O'Kelly G, Wu G et al. Reactome: A Database of Reactions, Pathways and Biological Processes, *Nucleic Acids Res.* 2011;39:D691-697.
281. Hubbard SJ, Thornton JM. NACCESS. London: Department of Biochemistry and Molecular Biology, University College, 1993.
282. Moffat JG, Rudolph J, Bailey D. Phenotypic Screening in Cancer Drug Discovery – Past, Present and Future, *Nat. Rev. Drug Discovery* 2014;13:588-602.
283. Swinney DC, Anthony J. How Were New Medicines Discovered?, *Nat. Rev. Drug Discovery* 2011;10:507-519.
284. Batash R, Asna N, Schaffer P et al. Glioblastoma Multiforme, Diagnosis and Treatment; Recent Literature Review, *Curr. Med. Chem.* 2017;24:3002-3009.
285. Stupp R, Mason WP, van den Bent MJ et al. Radiotherapy Plus Concomitant and Adjuvant Temozolomide for Glioblastoma, *N. Engl. J. Med.* 2005;352:987-996.
286. deCarvalho AC, Kim H, Poisson LM et al. Discordant Inheritance of Chromosomal and Extrachromosomal DNA Elements Contributes to Dynamic Disease Evolution in Glioblastoma, *Nat. Genet.* 2018;50:708-717.
287. Sottoriva A, Spiteri I, Piccirillo SG et al. Intratumor Heterogeneity in Human Glioblastoma Reflects Cancer Evolutionary Dynamics, *Proc. Natl. Acad. Sci. U. S. A.* 2013;110:4009-4014.
288. Finocchiaro G, Pellegatta S. Perspectives for Immunotherapy in Glioblastoma Treatment, *Curr. Opin. Oncol.* 2014;26:608-614.
289. Mangani D, Weller M, Roth P. The Network of Immunosuppressive Pathways in Glioblastoma, *Biochem. Pharmacol.* 2017;130:1-9.
290. Nduom EK, Weller M, Heimberger AB. Immunosuppressive Mechanisms in Glioblastoma, *Neuro-Oncology* 2015;17 Suppl 7:vii9-vii14.
291. Banissi C, Ghiringhelli F, Chen L et al. Treg Depletion with a Low-Dose Metronomic Temozolomide Regimen in a Rat Glioma Model, *Cancer Immunol. Immunother.* 2009;58:1627-1634.
292. Karachi A, Dastmalchi F, Mitchell D et al. Temozolomide for Immunomodulation in the Treatment of Glioblastoma, *Neuro-Oncology* 2018;20:1566–1572.
293. Litterman AJ, Zellmer DM, Grinnen KL et al. Profound Impairment of Adaptive Immune Responses by Alkylating Chemotherapy, *J. Immunol.* 2013;190:6259-6268.

294. Wu J, Waxman DJ. Metronomic Cyclophosphamide Eradicates Large Implanted GL261 Gliomas by Activating Antitumor Cd8(+) T-Cell Responses and Immune Memory, *OncoImmunology* 2015;4:e1005521.
295. Jones LH, Bunnage ME. Applications of Chemogenomic Library Screening in Drug Discovery, *Nat. Rev. Drug Discovery* 2017;16:285-296.
296. Wang Y, Cornett A, King FJ et al. Evidence-Based and Quantitative Prioritization of Tool Compounds in Phenotypic Drug Discovery, *Cell Chem. Biol.* 2016;23:862-874.
297. Arrowsmith CH, Audia JE, Austin C et al. The Promise and Peril of Chemical Probes, *Nat. Chem. Biol.* 2015;11:536-541.
298. Finan C, Gaulton A, Kruger FA et al. The Druggable Genome and Support for Target Identification and Validation in Drug Development, *Sci. Transl. Med.* 2017;9:eaag1166.
299. Irwin JJ, Gaskins G, Sterling T et al. Predicted Biological Activity of Purchasable Chemical Space, *J. Chem. Inf. Model.* 2018;58:148-164.
300. Wawer MJ, Li K, Gustafsdottir SM et al. Toward Performance-Diverse Small-Molecule Libraries for Cell-Based Phenotypic Screening Using Multiplexed High-Dimensional Profiling, *Proc. Natl. Acad. Sci. U. S. A.* 2014;111:10911-10916.
301. Edmondson R, Broglie JJ, Adcock AF et al. Three-Dimensional Cell Culture Systems and Their Applications in Drug Discovery and Cell-Based Biosensors, *Assay Drug Dev. Technol.* 2014;12:207-218.
302. Duval K, Grover H, Han LH et al. Modeling Physiological Events in 2D vs. 3D Cell Culture, *Physiology* 2017;32:266-277.
303. Dasari S, Tchounwou PB. Cisplatin in Cancer Therapy: Molecular Mechanisms of Action, *Eur. J. Pharmacol.* 2014;740:364-378.
304. Mitra A, Mishra L, Li S. Technologies for Deriving Primary Tumor Cells for Use in Personalized Cancer Therapy, *Trends Biotechnol.* 2013;31:347-354.
305. Maeda H, Khatami M. Analyses of Repeated Failures in Cancer Therapy for Solid Tumors: Poor Tumor-Selective Drug Delivery, Low Therapeutic Efficacy and Unsustainable Costs, *Clin. Transl. Med.* 2018;7:11.
306. Berman HM, Westbrook J, Feng Z et al. The Protein Data Bank, *Nucleic Acids Res.* 2000;28:235-242.
307. Xu D, Jalal SI, Sledge GW et al. Small-Molecule Binding Sites to Explore Protein-Protein Interactions in the Cancer Proteome, *Mol. BioSyst.* 2016;12:3067-3087.
308. Rolland T, Tasan M, Charlotheaux B et al. A Proteome-Scale Map of the Human Interactome Network, *Cell* 2014;159:1212-1226.

309. Li L, Wang B, Meroueh SO. Support Vector Regression Scoring of Receptor–Ligand Complexes for Rank-Ordering and Virtual Screening of Chemical Libraries, *J. Chem. Inf. Model.* 2011;51:2132-2138.
310. Wang H, Cai S, Bailey BJ et al. Combination Therapy in a Xenograft Model of Glioblastoma: Enhancement of the Antitumor Activity of Temozolomide by an MDM2 Antagonist, *J. Neurosurg.* 2017;126:446-459.
311. Herrera-Perez RM, Voytik-Harbin SL, Sarkaria JN et al. Presence of Stromal Cells in a Bioengineered Tumor Microenvironment Alters Glioblastoma Migration and Response to STAT3 Inhibition, *PLoS One* 2018;13:e0194183.
312. Sarkaria JN, Carlson BL, Schroeder MA et al. Use of an Orthotopic Xenograft Model for Assessing the Effect of Epidermal Growth Factor Receptor Amplification on Glioblastoma Radiation Response, *Clin. Cancer Res.* 2006;12:2264-2271.
313. Kang MH, Smith MA, Morton CL et al. National Cancer Institute Pediatric Preclinical Testing Program: Model Description for in vitro Cytotoxicity Testing, *Pediatr. Blood Cancer* 2011;56:239-249.
314. Bum-Erdene K, Zhou D, Gonzalez-Gutierrez G et al. Small-Molecule Covalent Modification of Conserved Cysteine Leads to Allosteric Inhibition of the TEAD-Yap Protein-Protein Interaction, *Cell Chem. Biol.* 2019;26:378-389.e313.
315. Kitange GJ, Carlson BL, Schroeder MA et al. Induction of MGMT Expression Is Associated With Temozolomide Resistance in Glioblastoma Xenografts, *Neuro-Oncology* 2009;11:281-291.
316. Cui Y, Brosnan JA, Blackford AL et al. Genetically Defined Subsets of Human Pancreatic Cancer Show Unique in Vitro Chemosensitivity, *Clin. Cancer Res.* 2012;18:6519-6530.
317. Jones S, Zhang X, Parsons DW et al. Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses, *Science* 2008;321:1801-1806.
318. Arpin CC, Mac S, Jiang Y et al. Applying Small Molecule Signal Transducer and Activator of Transcription-3 (STAT3) Protein Inhibitors as Pancreatic Cancer Therapeutics, *Mol. Cancer Ther.* 2016;15:794-805.
319. Folkman J. Tumor Angiogenesis: Therapeutic Implications, *N. Engl. J. Med.* 1971;285:1182-1186.
320. Arrillaga-Romany I, Reardon DA, Wen PY. Current Status of Antiangiogenic Therapies for Glioblastomas, *Expert Opin. Invest. Drugs* 2014;23:199-210.

321. Wang Y, Xing D, Zhao M et al. The Role of a Single Angiogenesis Inhibitor in the Treatment of Recurrent Glioblastoma Multiforme: A Meta-Analysis and Systematic Review, *PLoS One* 2016;11:e0152170.
322. Plate KH, Mennel HD. Vascular Morphology and Angiogenesis in Glial Tumors, *Exp. Toxicol. Pathol.* 1995;47:89-94.
323. Schmidt NO, Westphal M, Hagel C et al. Levels of Vascular Endothelial Growth Factor, Hepatocyte Growth Factor/Scatter Factor and Basic Fibroblast Growth Factor in Human Gliomas and Their Relation to Angiogenesis, *Int. J. Cancer* 1999;84:10-18.
324. Mi H, Muruganujan A, Casagrande JT et al. Large-Scale Gene Function Analysis with the PANTHER Classification System, *Nat. Protoc.* 2013;8:1551-1566.
325. Futreal PA, Coin L, Marshall M et al. A Census of Human Cancer Genes, *Nat. Rev. Cancer* 2004;4:177-183.
326. Wang Z, Yuan H, Sun C et al. GATA2 Promotes Glioma Progression Through EGFR/ERK/Elk-1 Pathway, *Med. Oncol.* 2015;32:87.
327. D'Alessio A, Proietti G, Lama G et al. Analysis of Angiogenesis Related Factors in Glioblastoma, Peritumoral Tissue and Their Derived Cancer Stem Cells, *Oncotarget* 2016;7:78541-78556.
328. Subramanian A, Narayan R, Corsello SM et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles, *Cell* 2017;171:1437-1452 e1417.
329. Neves SR, Ram PT, Iyengar R. G Protein Pathways, *Science* 2002;296:1636-1639.
330. Nilsson J, Sengupta J, Frank J et al. Regulation of Eukaryotic Translation by the RACK1 Protein: A Platform for Signalling Molecules on the Ribosome, *EMBO Rep.* 2004;5:1137-1141.
331. Lv QL, Huang YT, Wang GH et al. Overexpression of RACK1 Promotes Metastasis by Enhancing Epithelial-Mesenchymal Transition and Predicts Poor Prognosis in Human Glioma, *Int. J. Environ. Res. Public Health* 2016;13:1021.
332. Peng R, Jiang B, Ma J et al. Forced Downregulation of RACK1 Inhibits Glioma Development by Suppressing Src/Akt Signaling Activity, *Oncol. Rep.* 2013;30:2195-2202.
333. Kim JH, Lee SM, Lee JH et al. OGFOD1 Is Required for Breast Cancer Cell Proliferation and Is Associated With Poor Prognosis in Breast Cancer, *Oncotarget* 2015;6:19528-19541.
334. Okada Y, Sonoshita M, Kakizaki F et al. Amino-Terminal Enhancer of Split Gene AES Encodes a Tumor and Metastasis Suppressor of Prostate Cancer, *Cancer Sci.* 2017;108:744-752.
335. Vives V, Su J, Zhong S et al. ASPP2 Is a Haploinsufficient Tumor Suppressor That Cooperates With p53 to Suppress Tumor Growth, *Genes Dev.* 2006;20:1262-1267.

336. Schenone M, Dancik V, Wagner BK et al. Target Identification and Mechanism of Action in Chemical Biology and Drug Discovery, *Nat. Chem. Biol.* 2013;9:232-240.
337. Ong SE, Schenone M, Margolin AA et al. Identifying the Proteins to Which Small-Molecule Probes and Drugs Bind in Cells, *Proc. Natl. Acad. Sci. U. S. A.* 2009;106:4617-4622.
338. Nijman SM. Functional Genomics to Uncover Drug Mechanism of Action, *Nat. Chem. Biol.* 2015;11:942-948.
339. Lee J, Bogoy M. Target Deconvolution Techniques in Modern Phenotypic Profiling, *Curr. Opin. Chem. Biol.* 2013;17:118-126.
340. Kunimoto R, Dimova D, Bajorath J. Application of a New Scaffold Concept for Computational Target Deconvolution of Chemical Cancer Cell Line Screens, *ACS Omega* 2017;2:1463-1468.
341. Keiser MJ, Roth BL, Armbruster BN et al. Relating Protein Pharmacology by Ligand Chemistry, *Nat. Biotechnol.* 2007;25:197-206.
342. Li H, Gao Z, Kang L et al. TarFisDock: A Web Server for Identifying Drug Targets With Docking Approach, *Nucleic Acids Res.* 2006;34:W219-224.
343. Liu X, Baarsma HA, Thiam CH et al. Systematic Identification of Pharmacological Targets from Small-Molecule Phenotypic Screens, *Cell Chem. Biol.* 2016;23:1302-1313.
344. Savitski MM, Reinhard FB, Franken H et al. Tracking Cancer Drugs in Living Cells by Thermal Profiling of the Proteome, *Science* 2014;346:1255784.
345. Franken H, Mathieson T, Childs D et al. Thermal Proteome Profiling for Unbiased Identification of Direct and Indirect Drug Targets Using Multiplexed Quantitative Mass Spectrometry, *Nat. Protoc.* 2015;10:1567-1593.
346. Cox J, Mann M. MaxQuant Enables High Peptide Identification Rates, Individualized P.P.B.-Range Mass Accuracies and Proteome-Wide Protein Quantification, *Nat. Biotechnol.* 2008;26:1367-1372.
347. Szklarczyk D, Morris JH, Cook H et al. The STRING Database in 2017: Quality-Controlled Protein-Protein Association Networks, Made Broadly Accessible, *Nucleic Acids Res.* 2017;45:D362-D368.
348. Cianfrocco MA, DeSantis ME, Leschziner AE et al. Mechanism and Regulation of Cytoplasmic Dynein, *Annu. Rev. Cell Dev. Biol.* 2015;31:83-108.
349. Yu Y, Feng YM. The Role of Kinesin Family Proteins in Tumorigenesis and Progression: Potential Biomarkers and Molecular Targets for Cancer Therapy, *Cancer* 2010;116:5150-5160.

350. Kuramitsu Y, Suenaga S, Wang Y et al. Up-Regulation of DDX39 in Human Pancreatic Cancer Cells with Acquired Gemcitabine Resistance Compared to Gemcitabine-Sensitive Parental Cells, *Anticancer Res.* 2013;33:3133-3136.
351. Shi H, Cordin O, Minder CM et al. Crystal Structure of the Human ATP-Dependent Splicing and Export Factor UAP56, *Proc. Natl. Acad. Sci. U. S. A.* 2004;101:17628-17633.
352. Fayard E, Auwerx J, Schoonjans K. LRH-1: An Orphan Nuclear Receptor Involved in Development, Metabolism and Steroidogenesis, *Trends Cell Biol.* 2004;14:250-260.
353. Cai T, Liu Y, Xiao J. Long Noncoding RNA MALAT1 Knockdown Reverses Chemoresistance to Temozolomide via Promoting MicroRNA-101 in Glioblastoma, *Cancer Med.* 2018;7:1404-1415.
354. Eder J, Sedrani R, Wiesmann C. The Discovery of First-in-Class Drugs: Origins and Evolution, *Nat. Rev. Drug Discovery* 2014;13:577-587.
355. Moffat JG, Vincent F, Lee JA et al. Opportunities and Challenges in Phenotypic Drug Discovery: An Industry Perspective, *Nat. Rev. Drug Discovery* 2017;16:531-543.
356. Berns H, Humar R, Hengeler B et al. RACK1 Is Up-Regulated in Angiogenesis and Human Carcinomas, *FASEB J.* 2000;14:2549-2558.
357. Dave JM, Kang H, Abbey CA et al. Proteomic Profiling of Endothelial Invasion Revealed Receptor for Activated C Kinase 1 (RACK1) Complexed With Vimentin to Regulate Focal Adhesion Kinase (FAK), *J. Biol. Chem.* 2013;288:30720-30733.
358. Grossman RL, Heath AP, Ferretti V et al. Toward a Shared Vision for Cancer Genomic Data, *N. Engl. J. Med.* 2016;375:1109-1112.
359. Dobin A, Davis CA, Schlesinger F et al. STAR: Ultrafast Universal RNA-Seq Aligner, *Bioinformatics* 2013;29:15-21.
360. Anders S, Pyl PT, Huber W. HTSeq--A Python Framework to Work with High-Throughput Sequencing Data, *Bioinformatics* 2015;31:166-169.
361. Lun AT, Chen Y, Smyth GK. It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR, *Methods Mol. Biol.* 2016;1418:391-416.
362. Smedley D, Haider S, Durinck S et al. The BioMart Community Portal: An Innovative Alternative to Large, Centralized Data Repositories, *Nucleic Acids Res.* 2015;43:W589-598.
363. Cibulskis K, Lawrence MS, Carter SL et al. Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples, *Nat. Biotechnol.* 2013;31:213-219.
364. Shannon P, Markiel A, Ozier O et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks, *Genome Res.* 2003;13:2498-2504.

365. Sebaugh JL. Guidelines for Accurate EC50/IC50 Estimation, *Pharm. Stat.* 2011;10:128-134.
366. Basavarajappa HD, Sulaiman RS, Qi X et al. Ferrochelatase Is a Therapeutic Target for Ocular Neovascularization, *EMBO Mol. Med.* 2017;9:786-801.
367. Carpentier G. Angiogenesis Analyzer for ImageJ. 2012.
368. Liao Y, Smyth GK, Shi W. featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features, *Bioinformatics* 2014;30:923-930.
369. McCarthy DJ, Chen Y, Smyth GK. Differential Expression Analysis of Multifactor RNA-seq Experiments with Respect to Biological Variation, *Nucleic Acids Res.* 2012;40:4288-4297.
370. Dennis G, Jr., Sherman BT, Hosack DA et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery, *Genome Biol.* 2003;4:R60.
371. Huang da W, Sherman BT, Lempicki RA. Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources, *Nat. Protoc.* 2009;4:44-57.
372. Connelly KE, Hedrick V, Paschoal Sobreira TJ et al. Analysis of Human Nuclear Protein Complexes by Quantitative Mass Spectrometry Profiling, *Proteomics* 2018;18:e1700427.
373. Schwanhaussner B, Busse D, Li N et al. Global Quantification of Mammalian Gene Expression Control, *Nature* 2011;473:337-342.
374. Jafari R, Almqvist H, Axelsson H et al. The Cellular Thermal Shift Assay for Evaluating Drug Target Interactions in Cells, *Nat. Protoc.* 2014;9:2100-2122.
375. Si Y, Xu D, Bum-Erdene K et al. Chemical Space Overlap with Critical Protein-Protein Interface Residues in Commercial and Specialized Small-Molecule Libraries, *ChemMedChem* 2019;14:119-131.

CURRICULUM VITAE

David Xu

Education

- 2019 Ph.D., Bioinformatics with a Minor in Computer Science Indiana University
- 2018 M.S., Bioinformatics Indiana University
- 2010 B.S., Biology with a Minor in Chemistry University of Illinois

Peer-Reviewed Publications

- Bum-Erdene, K.; Zhou, D.; Gonzalez-Gutierrez, G.; Ghozayel, M.; Si, Y.; **Xu, D.**; Shannon, H.E.; Bailey, B. J.; Corson, T. W.; Pollok, K. E.; Wells, C. D.; Meroueh, S. O., Small-Molecule Covalent Modification of Conserved Cysteine Leads to Allosteric Inhibition of the TEAD4•Yap1 Protein-Protein Interaction. *Cell Chem Biol* **2019**, 26 (3), 378-389.e13.
- Si, Y.; **Xu, D.**; Bum-Erdene, K.; Ghozayel, M.; Yang, B.; Clemons, P. A.; Meroueh, S. O., Chemical Space Overlap with Critical Protein-Protein Interface Residues in Commercial and Specialized Small-Molecule Libraries. *ChemMedChem* **2019**, 14 (1), 119-131.
- Zhou, D.; Bum-Erdene, K.; **Xu, D.**; Liu, D.; Tompkins, D.; Sulaiman, R.S.; Corson, T.W.; Chirgwin, J. M.; Meroueh, S. O., Small Molecules Inhibit Ex Vivo Tumor Growth in Bone. *Bioorg Med Chem* **2018**, 26 (23-24), 6128-6134.
- Chen, X.; Liu, D.; Zhou, D.; Si, Y.; **Xu, D.**; Stamatkin, C. W.; Ghozayel, M.; Ripsch, M. S.; Obukhov, A. G.; White, F. A.; Meroueh, S. O., Small-Molecule $Ca_v\alpha_1$ • $Ca_v\beta$ Antagonist Suppresses Neuronal Voltage-Gated Calcium Channel Trafficking. *Proc Natl Acad Sci* **2018**, 115 (45), E10566-E10575.
- **Xu, D.**; Bum-Erdene, K.; Si, Y.; Zhou, D.; Ghozayel, M.; Meroueh, S. O., Mimicking Intermolecular Interactions of Tight Protein-Protein Complexes for Small-Molecule. *ChemMedChem* **2017**, 12 (21), 1794-1809.
- **Xu, D.**; Si, Y.; Meroueh, S. O., A Computational Investigation of Small-Molecule Engagement of Hot Spots at Protein-Protein Interactions Interfaces. *J Chem Inf Model* **2017**, 57, 2250-2272.
- **Xu, D.**; Li, L.; Zhou, D.; Liu, D.; Hudmon, A.; Meroueh, S. O., Structure-Based Target-Specific Screening Leads to Small-Molecules CaMKII Inhibitors. *ChemMedChem* **2017**, 12, 660-677.
- Zhou, D.; Maya, Z.; **Xu, D.**; Liu, D.; Hudmon, A.; Macleod, K. F.; Meroueh, S. O., Small Molecules Inhibit STAT3 Activation, Autophagy, and Cancer Cell Anchorage-Independent Growth. *Bioorg Med Chem* **2017**, 25, 2995-3005.

- Liu, D.*; **Xu, D.***; Liu, M.; Knabe, E. W.; Yuan, C.; Huang, M.; Meroueh, S. O., Small Molecules Engage Hot Spots Through Cooperative Binding to Inhibit a Tight Protein-Protein Interaction. *Biochemistry* **2017**, 56, 1768-1784. (**Co-First Author*)
- **Xu, D.**; Jalal, S.; Sledge, G. W.; Meroueh, S. O., Small-Molecule Binding Sites to Explore New Protein-Protein Interactions in the Cancer Proteome. *Mol Biosyst* **2016**, 12, 3067-3087.
- **Xu, D.**; Meroueh, S. O., Effect of Binding Pose and Modeled Structures on SVMGen and GlideScore Enrichment of Chemical Libraries. *J Chem Inf Model* **2016**, 56, 1139-1151.
- **Xu, D.**; Wang, B.; Meroueh, S. O., Structure-Based Computational Approaches for Small-Molecule Modulation of Protein-Protein Interactions. *Methods Mol Biol* **2015**, 1278, 77-92.
- Peng, X., Wang, F.; Li, L.; Bum-Erdene, K.; **Xu, D.**; Wang, B.; Sinn, A. A.; Pollok, K. E.; Sandusky, G. E.; Li, L.; Turchi, J. J.; Jalal, S. I.; Meroueh, S. O., Exploring a Structural Protein-Drug Interactome for New Therapeutics in Lung Cancer. *Mol Biosyst* **2014**, 10, 581-591.

Conference Papers

- Quick, R.; Hayashi, S.; Meroueh, S. O.; Rynge, M.; Teige, S.; Wang, B.; **Xu, D.**, Building a Chemical-Protein Interactome on the Open Science Grid. *International Symposium on Grids and Clouds (ISGC) 2015*, Taipei, Taiwan. March 15-20, 2015.
- Tavares, M.; Quick, R.; Teige, S.; Rynge, M.; Meroueh, S.O.; **Xu, D.**; Wang, B.; Hayashi, S., Building a Chemical-Protein Interactome on the Open Science Grid. *High Performance Parallel and Distributed Computer '14*, Vancouver, Canada, June 23-27, 2014.

Poster Presentations

- **Xu, D.**; Zhou, D.; Bum-Erdene, K.; Bailey, B. J.; Liu, S.; Wan, J.; Fishel, M. L.; Aryal, U. K.; Lee, J. A.; Corson, T. W.; Pollok, K. E.; Meroueh, S. O., Tumor-Specific Chemogenomic Libraries by Structure-Based Enrichment for Glioblastoma Phenotypic Screening. *IU Simon Cancer Center's Cancer Research Day*, Indianapolis, IN, May 15, 2019.
- **Xu, D.**; Si, Y.; Meroueh, S. O., A Computational Investigation of Small-Molecule Engagement of Hot Spots at Protein-Protein Interactions Interfaces. *Intelligent Systems for Molecular Biology (ISMB) 2018*, Chicago, IL, July 6-10, 2018.
- **Xu, D.**; Liu, D.; Ghozayel, M.; Joshi, N.; Meroueh, S.O., A Computational Screen Guided by Protein Interface Hot Spots Leads to Small-Molecule Antagonists. *Statewide Structural Biology Forum*, Indianapolis, IN, March 2, 2016.
- **Xu, D.**; Meroueh, S. O., Druggable Proteome of The Cancer Genome Atlas (TCGA). *IU Simon Cancer Center's Cancer Research Day*, Indianapolis, IN, May 29, 2014.