

MODELLING AND DISRUPTING PROTEIN INTERACTIONS

By

Nicolas Arcenio Pabon

B.S. in Physics, Carnegie Mellon University, 2013

Submitted to the Graduate Faculty of

the School of Medicine in partial fulfillment

of the requirements for the degree of

Ph.D. in Computational Biology

University of Pittsburgh
2018

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Nicolas Arcenio Pabon

It was defended on

July 18, 2018

and approved by

Dr. Ivet Bahar, Ph.D., Distinguished Professor & JK Vries Chair, Computational and Systems Biology
Department, University of Pittsburgh School of Medicine

Dr. Ziv Bar-Joseph, Ph.D., FORE Systems Professor, Machine Learning and Computational Biology
Departments, Carnegie Mellon University

Dr. Neil Hukriede, Ph.D., Vice Chair, Department of Developmental Biology, University of Pittsburgh
School of Medicine

Dissertation Advisor: Dr. Carlos J. Camacho, Ph.D., Associate Professor, Computational and Systems
Biology Department, University of Pittsburgh School of Medicine

Copyright © by Nicolas Arcenio Pabon
2018

MODELLING AND DISRUPTING PROTEIN INTERACTIONS

Nicolas Arcenio Pabon, Ph.D.

University of Pittsburgh, 2018

Rational drug design requires a deep understanding of protein interactions, both in terms of the structural mechanisms that regulate binding of individual molecules and the broader, pathway- and cell-level effects of disrupting protein interaction networks. This thesis presents work that stems from these ideas, discussing contributions to a number of current challenges in the field of drug discovery. First, we describe how structural flexibility is leveraged by ‘selectively promiscuous’ protein interfaces – enabling them to bind specifically to several distinctly shaped ligands. Taking PD-1 as a case study, we demonstrate using molecular dynamics simulations how the flexible receptor interface recognizes conserved ‘trigger’ motifs on its cognate ligands’ interfaces. Trigger interactions, which we show are also exploited by a recent blockbuster PD-1 inhibitor, drive the initial steps of an induced-fit binding pathway that then ‘splits’ into distinct, ligand-specific bound states. Second, we present a hybrid genomic and structural pipeline for genome-scale identification of protein targets for bioactive compounds. We train a random forest classifier to predict compound-target interactions from compound treatment and gene knockdown gene expression signatures in multiple cell types. Refining genomic predictions with a structure-based screen, we achieve top-10/top-100 target prediction accuracies of 26%/41%, respectively, on a validation set of 152 FDA-approved drugs, and validate previously unknown small molecule modulators of HRAS, KRAS, CHIP, and PDK1. Third, we present a strategy that combines transcriptomic and structural analyses with single-cell microscopy to predict small molecule inhibitors of TNF-induced NF- κ B signaling and elucidate

the network response. Validating two novel pathway inhibitors that disrupt the protein network upstream of IKK and NF- κ B, our findings suggest that a network-centric drug discovery approach is a promising strategy to evaluate the impact of pharmacologic intervention in signaling. Last, we introduce DrugQuery (DQ), a structure-based public web server for small molecule target prediction. DQ docks user-submitted small molecules against a target library of 7957 predicted binding sites on 1245 human proteins. The server achieved a top-decile target prediction accuracy of 68% on a validation set of 95 FDA-approved drugs and 86% on a validation set of 102 FXR-binding compounds from the 2017 D3R Grand Challenge 2.

TABLE OF CONTENTS

1.0 INTRODUCTION	1
1.1. OUTLINE	1
2.0 PROBING PROTEIN FLEXIBILITY REVEALS A MECHANISM FOR SELECTIVE PROMISCUITY	4
2.1. INTRODUCTION	4
2.2. RESULTS	11
2.2.1. Open and closed states of PD-1 Asn66 and Ile126 describe a hydrophilic or hydrophobic interface	11
2.2.2. Bound-like conformations of unbound Tyr123/112 in PD-L1/2 facilitates molecular recognition	14
2.2.3. Conserved PD-L1/2 Asp122/111 form a specific intermolecular hydrogen bond network that opens PD-1 Asn66 and switches the receptor interface from hydrophilic to hydrophobic	15
2.2.4. ADY/GDY ligand motifs stabilize distinct bound-like states for PD-L1/2	17
2.2.5. PD-L1/2 triggers ADY/GDY produce energetically downhill induced fit binding pathways	21
2.2.6. Encounter complex simulations suggest chronology of induced fit triggering interactions	22
2.2.7. PD-1 – targeting antibody validates the critical role of Asn66 and suggests an anchor-independent binding mechanism with closed Ile126 and Ile134	25

2.2.8. Can molecular triggers be exploited to drug PD-1?	28
2.3. DISCUSSION	31
2.3.1. Induced fit motif XDY shared by PD-1 ligands modulates the flexible PD-1 binding interface from hydrophilic to hydrophobic	31
2.3.2. A single carbon atom difference can shift the hydrophobic PD-1 binding surface from a stable patch to a stable cavity	32
2.3.3. Bound-like XDY residues and molecular recognition	32
2.3.4. Downhill binding pathways strongly suggest an induced fit binding mechanism	33
2.3.5. Two step binding pathway of PD-1 reveals a simple mechanism for selective promiscuity	34
2.3.6. Molecular triggers could be exploited to design small-molecule PD-1 antagonists	35
2.3.7. Selective promiscuity via induced fit offers potential advantages over conformational selection for multi-ligand regulatory proteins	35
2.4. METHODS	36
2.4.1. Initial protein structures used in simulations	36
2.4.2. Peptide ligand mimics used in simulations	38
2.4.3. Simulating PD-1 – peptide interactions	38
2.4.4. Encounter complex modeling and simulation	39
2.4.5. Simulation parameters	40
2.4.6. Analysis tools	40
2.4.7. Relative free energies of bound-like versus non-bound-like interfaces	41
2.5. SUPPLEMENTARY FIGURES	43
3.0 PREDICTING PROTEIN TARGETS FOR DRUG-LIKE COMPOUNDS USING TRANSCRIPTOMICS	50
3.1. INTRODUCTION	50

3.2. RESULTS	54
3.2.1. Preliminary prediction of drug targets using expression profile correlation features	54
3.2.2. Combining individual features using random forest	57
3.2.3. Gene ontology analysis of protein targets	59
3.2.4. Structural enrichment of genomic predictions	59
3.2.5. Identifying new interactions in the LINCS dataset	61
3.2.6. Target-centric prediction of novel RAS inhibitors	63
3.2.7. Target-centric prediction of novel CHIP inhibitors	64
3.2.8. Compound-centric prediction of a novel target for the drug Wortmannin	68
3.2.9. Comparison to existing target prediction method	70
3.3. DISCUSSION	71
3.4. METHODS	74
3.4.1. Data sources	74
3.4.2. Extracting experiments from LINCS	75
3.4.3. Building a validation dataset from LINCS	77
3.4.4. Extracting and integrating features from different data sources	78
3.4.5. Subcellular localization assignment	81
3.4.6. Classification procedure	82
3.4.7. Extending random forests to drugs with missing features	83
3.4.8. Generating structural models for docking	84
3.4.9. Docking procedure	86
3.4.10. Comparison to previous expression perturbation target prediction methods	86
3.4.11. Experimental Assays involving HRAS and KRAS	86
3.4.12. Experimental Assays involving CHIP	87

3.4.13. Experimental assays involving PDK1	90
3.5. SUPPLEMENTARY FIGURES	92
3.6. SUPPLEMENTARY TABLES	96
3.7. ADDITIONAL FILES	100
4.0 DRUGGING THE TNF-INDUCED NF-κB SIGNALING NETWORK	101
4.1. INTRODUCTION	101
4.2. RESULTS	104
4.3. DISCUSSION	109
4.4. METHODS	110
4.4.1. Analysis of gene expression data	110
4.4.2. Structural analysis	111
4.4.3. Thermal shift assay and analysis	112
4.4.4. Establishing EGFP- RELA /IKK γ CRISPR Knock-in Cells	112
4.4.5. Western blot analysis	114
4.4.6. Live-cell imaging and analysis	114
4.4.7. Fixed-cell immunofluorescence and analysis	115
4.5. SUPPLEMENTARY FIGURES	117
5.0 DRUGQUERY: STRUCTURE-BASED SMALL MOLECULE VIRTUAL SCREENING OF HUMAN PROTEINS	
.....	122
5.1. INTRODUCTION	122
5.2. METHODS	124
5.2.1. The DrugQuery target library	124
5.2.2. Docking and scoring	125
5.2.3. RMSD analysis of predicted binding modes	126

5.2.4. Web server	126
5.3. USING DRUGQUERY	127
5.3.1. Uploading a new small molecule	127
5.3.2. Tracking a job	129
5.3.3. Downloading results	129
5.4. VALIDATION	129
5.4.1. Predicting targets of FDA-approved drugs	129
5.4.2. Predicting targets of non-drug bioactive compounds	132
5.5. SUMMARY	135
5.6. ADDITIONAL FILES	136
6.0 CONCLUSIONS AND FUTURE RESEARCH	137
7.0 BIBIOGRAPHY	140

LIST OF TABLES

Table 2.1. Anchor Tyr123 is key determinant of bound-like docked conformations	15
Table 2.2. Free energy difference between the non-bound-like and bound-like states of PD-1 interface residues Asn66 and Ile126 in various systems	20
Table 2.3. Chronology of the formation of intermolecular interactions between PD-1 and PD-L1/2 in encounter complex simulations	23
Table 2.4. Top 5 PD-1 residues contributing to electrostatic energy when binding to PD-L1 and pembrolizumab	27
Table 2.5. Top 5 PD-1 residues contributing to desolvation energy when binding to PD-L1 and pembrolizumab	27
Table 3.1. Performance of target prediction using different features and methods on the 29 FDA-approved drugs tested in 7 cell lines	56
Table 3.2. Performance of two random forest models on validation set of 152 FDA-approved drugs as a function of cells tested	58
Table 3.3. Comparison of our pipeline to existing drug-target prediction methods	71
Table 3.4. The cellular localization of successful and unsuccessful drug targets enriched by gene ontology	96
Table 3.5. Predicted HRAS/KRAS-targeting compounds purchased for experimental validation	97
Table 3.6. Predicted CHIP-targeting compounds purchased for experimental testing	97
Table 3.7. Symbols and notations	98

Table 3.8. Summary of constructed feature sets	99
Table 3.9. Cell lines included in the validation dataset	99

LIST OF FIGURES

Figure 2.1 General mechanism for ligand binding to flexible receptor	6
Figure 2.2 Flexibility of the PD-1 binding interface	8
Figure 2.3. Structures of PD-L1/2 – mimicking peptides used to probe PD-1 interface dynamics	10
Figure 2.4. Dynamics of PD-1 binding interface in the presence of different ligands	12
Figure 2.5. Hydrogen bond network of PD-1 Asn66 in different contexts	13
Figure 2.6. Stabilization of bound-like Ile134 by conserved tyrosine (Y) anchor	18
Figure 2.7. Downhill binding pathways of PD-1 triggers of induced fit for each cognate ligand	19
Figure 2.8. Modulation of the PD-1 interface binding cavity in encounter complex simulations with PD-L1 and PD-L2	24
Figure 2.9. Secondary, non-triggering contacts in PD-1 encounter complexes	24
Figure 2.10. Pembrolizumab – bound PD-1 interface resembles PD-L1 – bound interface with a closed Ile134	26
Figure 2.11. Macrocyclic mGDV motif induces structural changes in the PD-1 interface towards the pembrolizumab – bound state	30
Figure 2.12: Stability of apo PD-1 simulations	37
Figure 2.13. The cognate ligands of PD-1	43
Figure 2.14. Modulation of PD-1's flexible interface cavity	44
Figure 2.15. Apo PD-1 interactions with GDY peptide opens a hydrophobic cavity	45
Figure 2.16. Replicate trajectories from Figure 2.4a,b,c	46

Figure 2.17. Dynamics of PD-1 binding cavity in the presence of different anchor substitutes	47
Figure 2.18. Model of potent Bristol-Myers-Squibb macrocyclic PD-1 inhibitor	48
Figure 2.19. Predicted interactions of Bristol-Myers-Squibb macrocyclic PD-1 inhibitor	49
Figure 3.1. Drug and gene knockdown induced mRNA expression profile correlations reveal drug-target interactions	53
Figure 3.2. Structural enrichment of genomic target predictions	61
Figure 3.3. Workflow of combined genomic (green) and structural (blue) pipeline for drug-target interaction prediction	62
Figure 3.4. HRAS/KRAS inhibitors predicted based on direct correlations and docked poses show direct binding in SPR assays	63
Figure 3.5. Predicted inhibitors show direct binding to and functional inhibition of CHIP	66
Figure 3.6. mRNA expression signature of CHIP inhibitor 2.1 correlates with knockdown of CHIP interacting partners	67
Figure 3.7. Wortmannin promotes PDK1 – PIP3 binding in vitro	69
Figure 3.8. Comparing random forest approaches with a random classifier for predicting known targets of validation compounds	92
Figure 3.9. ZINC compounds weakly disrupt CHIP binding to chaperone peptide as measured by fluorescence polarization	92
Figure 3.10. CHIP inhibitors prevent ubiquitination by CHIP in vitro	93
Figure 3.11. Predicted CHIP inhibitors prevent ubiquitination of an alternate substrate	94
Figure 3.12. Comparison of gene expression-based and pharmacophore-based virtual screens against CHIP	95
Figure 4.1. Small molecule treatments produce transcriptional responses in that correlate with genetic knockdowns of proteins involved in NF- κ B signaling	103

Figure 4.2. Thermal shift assays indicate moderate dose-dependent stabilization of TRAF2 by compounds 2 and 3	105
Figure 4.3. Small molecules disruptors of NF- κ B signaling reduce nuclear translocation of NF- κ B and the formation of NEMO puncta in TNF-stimulated cells	108
Figure 4.4. Prediction pipeline used to identify small molecule inhibitors of TNF-inducible NF- κ B signaling	117
Figure 4.5. Predicted binding mode of compound 1 to TRADD-binding interface of TRAF2	118
Figure 4.6. Thermal shift assays indicate no clear effect of compound 1 on TRAF2 stability	118
Figure 4.7. Quantification of FP-RelA expression in U2OS cells	119
Figure 4.8. Other descriptors of nuclear FP-RelA	119
Figure 4.9. Compound 1 does not have a significant effect on FP-RelA translocation	120
Figure 4.10. Western blot of IKKy	120
Figure 4.11. IKKy expression in the presence of compounds 2 and 3	121
Figure 5.1. Profiles of protein structures and predicted binding sites in the DrugQuery library	125
Figure 5.2. The DrugQuery user interface	128
Figure 5.3. Properties of FDA-approved compounds in Validation Set 1	130
Figure 5.4. DrugQuery target prediction accuracy on Validation Set 1	131
Figure 5.5. DrugQuery predicts native-like poses for compounds in Validation Set 1	132
Figure 5.6. Properties of compounds in Validation Set 2 and DrugQuery target prediction accuracy ...	133
Figure 5.7. Native-like predicted poses for compounds in Validation Set 2	134
Figure 5.8. Unique receptor conformations in Validation Set 2 produce “swapped” hydrophobic contacts in predicted poses	134

1.0 INTRODUCTION

Rational drug design is, in essence, the effort to develop leads in drug discovery by leveraging *all known theoretical and experimental knowledge* of the system one is trying to drug [1]. Owing to decades of progress in fields such as computer science, statistics, machine learning, genomics, molecular biology, and biochemistry, rational drug design has obtained a central role in medicinal chemistry as a more cost- and time-efficient complement to traditional high-throughput screening [2]. Despite significant technological and scientific advances, however, global attrition rates in pharmaceutical programs are climbing, as is the average R&D cost per drug approved [3-5]. We currently face important challenges in two key areas of rational drug design: structure-based methods [6], which attempt to leverage the 3D structures of protein disease targets to discover small molecule (ant)agonists, and genomic methods [7], which attempt to leverage gene expression and other multi-omic data to identify disease targets and predict potential inhibitors. These challenges include: (1) understanding how the flexibility of protein interfaces regulates specificity and promiscuity to binding partners, (2) predicting the protein targets of bioactive small molecules, (3) developing strategies for small molecule modulation of complex signaling networks, and (4) large-scale structure-based compound-centric target screening. In this dissertation we will discuss our advances in these four areas.

1.1 OUTLINE

This dissertation is organized as follows:

In Chapter 2.0 we discuss how the human PD-1 receptor exploits interface flexibility to achieve “selectively promiscuous” binding to structurally-distinct cognate ligands. We use molecular dynamics simulations to identify the mechanisms that trigger structural transitions between the unbound and bound PD-1 interface conformations. Our results show that conserved “triggers” on the ligand interfaces drive the initial steps of an induced-fit binding pathway that then splits into two distinct bound states, with PD-1’s flexibility accommodating the non-conserved, trigger-adjacent ligand features. We demonstrate that PD-1’s ‘selective promiscuity’ results largely from the displacement of its interface residue Asn66 by trigger interactions, which switches the receptor interface from flat and polar to either a hydrophobic patch or a hydrophobic cavity, depending on which ligand drives the induced fit transition. A recently published crystal structure of Pembrolizumab, a blockbuster PD-1 - targeting immune checkpoint inhibitor, confirms the importance of conserved triggering interactions in binding to flexible protein interfaces.

In Chapter 3.0 we demonstrate a hybrid genomic & structural target prediction pipeline for bioactive small molecules. Using gene expression data from thousands of small molecule treatment and gene knockdown experiments in live cells, we train a random forest classifier to predict compound-protein interactions from correlations between compound & knockdown expression signatures. We then refine our genomic target predictions with structural modeling and molecular docking. On a validation set of 152 FDA-approved drugs and 3104 potential targets we achieve top-10 and top-100 target prediction accuracies of 26% and 41%, respectively, doubling the accuracy of previous gene expression-based methods. We additionally validate several previously unknown small molecule modulators of HRAS, KRAS, CHIP, and PDK1.

In Chapter 4.0 we discuss an application of the insight we gained from Chapter 3.0 – an extension of our pipeline that combines its transcriptomic and structural analysis with single-cell microscopy to predict small molecule disruptors of TNF-induced NF- κ B signaling and characterize the response of the disrupted protein interaction network. Using live-cell fluorescence assays to monitor signaling dynamics of cells treated with predicted disruptors of the canonical NF- κ B pathway, we identify two compounds that inhibit formation of the mature TNFR1 complex, preventing recruitment of the IKK complex and eliminating NF- κ B translocation to the nucleus.

In Chapter 5.0 we present a public web server that we developed for molecular docking-based small molecule target fishing. The server, called DrugQuery (DQ), docks user-uploaded small molecules against 1245 human proteins and returns ranked target predictions and structural models of predicted binding modes. Approximately 8000 precomputed binding sites across the DQ target library are stored in a database and used to accelerate docking, which enables results to be returned in mere hours for most small molecules. On a validation set of 95 FDA-approved with known target structures, DQ correctly predicts the known target in the top decile of potential targets 68% of the time. On a separate validation set of 102 congeneric FXR-binding compounds from the 2017 D3R Grand Challenge 2, DQ achieves a 86% top decile prediction accuracy.

2.0 PROBING PROTEIN FLEXIBILITY REVEALS A MECHANISM FOR SELECTIVE PROMISCUITY

2.1 INTRODUCTION

Structural and proteomic research over the past decade has supplanted the traditional structure-function paradigm by establishing the functional relevance of protein dynamics [8-13]. In particular, eukaryotic regulatory and signaling proteins are skewed towards notably higher degrees of flexibility when compared to other functional categories [14, 15]. Regulatory proteins also tend towards comparatively higher degrees of binding promiscuity, and we have previously shown thermodynamically how the entropy associated with their flexibility can relate to their specificity towards multiple binding partners [14]. However, a structural understanding of how this selective promiscuity is achieved is still lacking.

Flexible human regulatory proteins such as MDM2 and PD-1 usually only crystallize when ligand-bound. Although Nuclear Magnetic Resonance (NMR) can occasionally resolve unbound (apo) structures of these proteins, it is noteworthy that their apo NMR ensembles often deviate from their bound crystal structures [16-22]. Thus, for many such proteins, available structural data do not capture the full binding dynamics, and the pathway from the apo, non-bound-like state to the bound-like state is unclear. This lack of data obscures the mechanistic connection between interface flexibility, binding promiscuity, and ligand specificity. Moreover, given that many regulatory proteins are promising drug targets, this missing puzzle piece often spells failure for drug design efforts that only target the bound-like state, assuming that this state is available in the apo ensemble. Rational approaches to target flexible proteins will thus benefit

from new methods that can reveal the binding pathways connecting the non-bound-like to the bound-like states.

Binding to flexible receptors is traditionally described by conformational selection [23, 24] or induced fit [25] mechanisms, and NMR techniques are often used to distinguish between these two (Figure 2.1). Generally speaking, one assumes a conformational selection scenario if the apo protein ensemble samples bound-like states (apo_{BL}) [26, 27]. If not, one assumes induced fit [16]. In reality, whether a protein-protein interaction occurs via conformational selection or induced fit depends on the flux of the system through the two alternate pathways from the non-bound-like apo state (apo_{NBL}) to the bound-like encounter complex (EC_{BL}) [28]. Flux through the conformational selection pathway is limited by the free energy difference between the apo_{BL} and apo_{NBL} states, $\Delta G_{\text{BL}}^{\text{apo}}$, which determines the fractional population of the bound-like state and thus restricts when selection-association with the ligand can occur. On the other hand, flux through the induced fit pathway is for the most part independent of $\Delta G_{\text{BL}}^{\text{apo}}$, as the ligand is presumed to be able to associate with all apo receptor microstates. Instead, flux through this pathway is limited by the free energy difference between the EC_{BL} and the non-bound-like encounter complex (EC_{NBL}), $\Delta G_{\text{BL}}^{\text{EC}}$, which is a function of specific interactions between receptor and ligand, and the energy barrier between these states. Both pathways terminate via a ubiquitous optimization step in which minor structural rearrangements at the EC_{BL} interface lead to the high affinity complex.

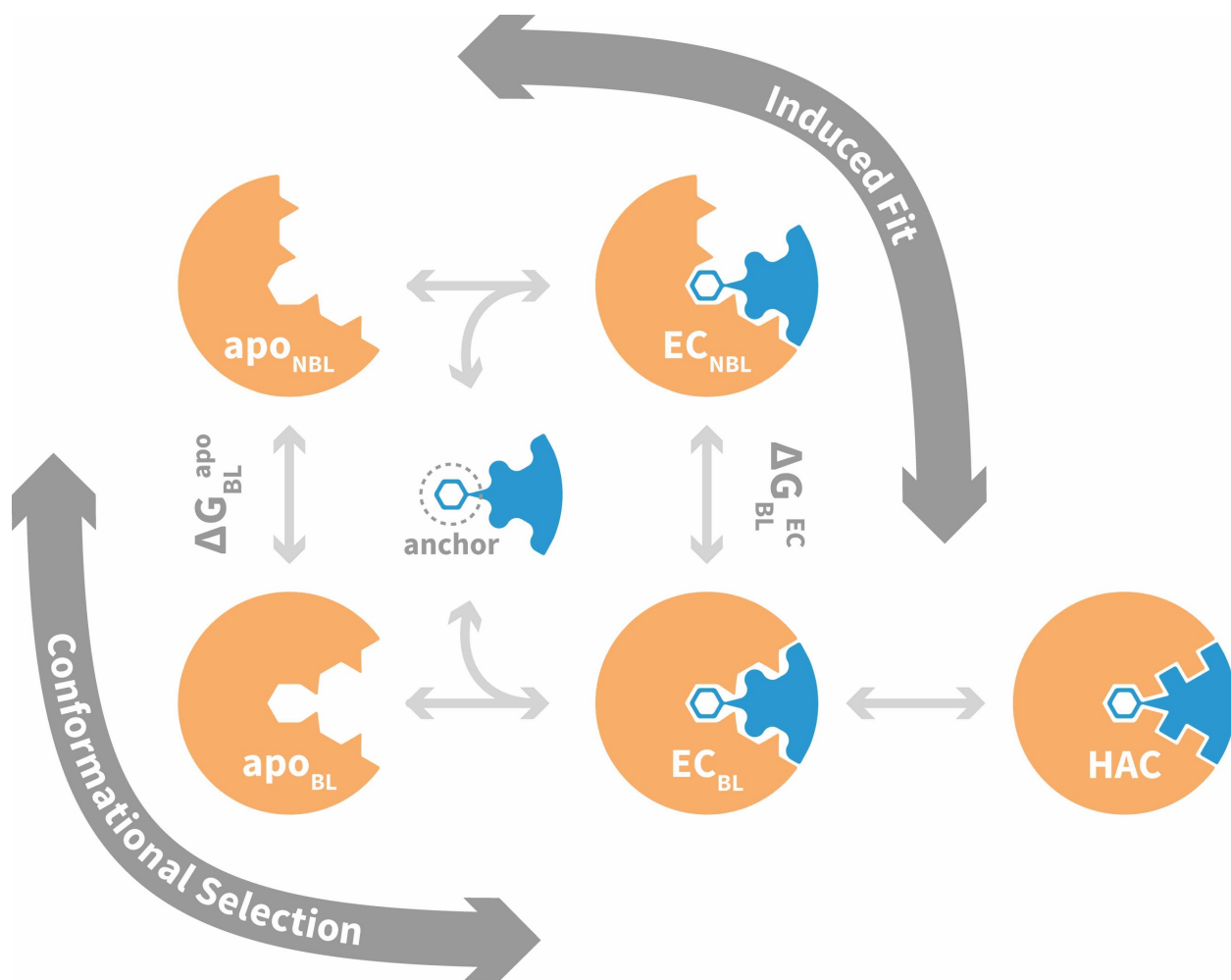


Figure 2.1. General mechanism for ligand binding to flexible receptor. In the conformational selection pathway, the ligand docks to the bound-like (BL) form of the apo receptor (apo_{BL}) to form the bound-like encounter complex (EC_{BL}). In the induced fit pathway, the ligand docks to the non-bound-like (NBL) form of the apo receptor (apo_{NBL}) to form the non-bound-like encounter complex (EC_{NBL}). Intermolecular interactions then drive structural transitions to the EC_{BL}. Both pathways end with a final induced fit step that optimizes interface side chains, transitioning to the high affinity complex (HAC). The binding mechanism also highlights an anchor residue often found to be important in molecular recognition [29].

To shed light on the structural basis of selective promiscuity in the aforementioned class of flexible-interface multi-ligand proteins, we study the binding mechanism of PD-1 to its cognate ligands PD-L1 and PD-L2. Human PD-1 is a T cell receptor and immune response regulator that has recently emerged as a breakthrough anti-cancer target [30, 31]. NMR and crystallographic studies have revealed the flexibility of the PD-1 interface by showing that its apo and bound conformations are very different [19-22] (Figure 2.2, Figure 2.13), suggestive of an induced fit mechanism. Specifically, while the apo PD-1 interface shows a polar surface around Asn66 with an unmatched NH₂ (Figure 2.2a), in complex this NH₂ group forms two hydrogen bonds, with the PD-L1 – bound interface exhibiting a hydrophobic patch around Ile126 (Figure 2.2b), and the PD-L2 – bound interface forming a large hydrophobic cavity flanked by Ile126 and Ile134 (Figure 2.2c, Figure 2.14).

To date, no small molecular weight PD-1 inhibitors have been reported in the literature despite the importance of this blockbuster target [30-32]. This was somewhat surprising, since the Trp110 binding site observed in the PD-L2 – bound cocrystal (Figure 2.2c) displays two key characteristics known to be favorable for ligand binding: concavity [33, 34] and hydrophobicity [35]. It is reasonable to assume that the flexibility of the Trp110 pocket, and the fact that in the apo state it is largely occluded by the unmatched, polar NH₂ group of Asn66 (Figure 2.2a,d), would present significant obstacles to traditional structure-based drug-design methods attempting to target this cavity [36]. Thus, efforts to model the binding mechanism of PD-1 would not only shed light on nature's design principles for flexible and promiscuous protein-protein interfaces, but they may also offer novel avenues for pursuing rational drug design against this and other high-impact targets.

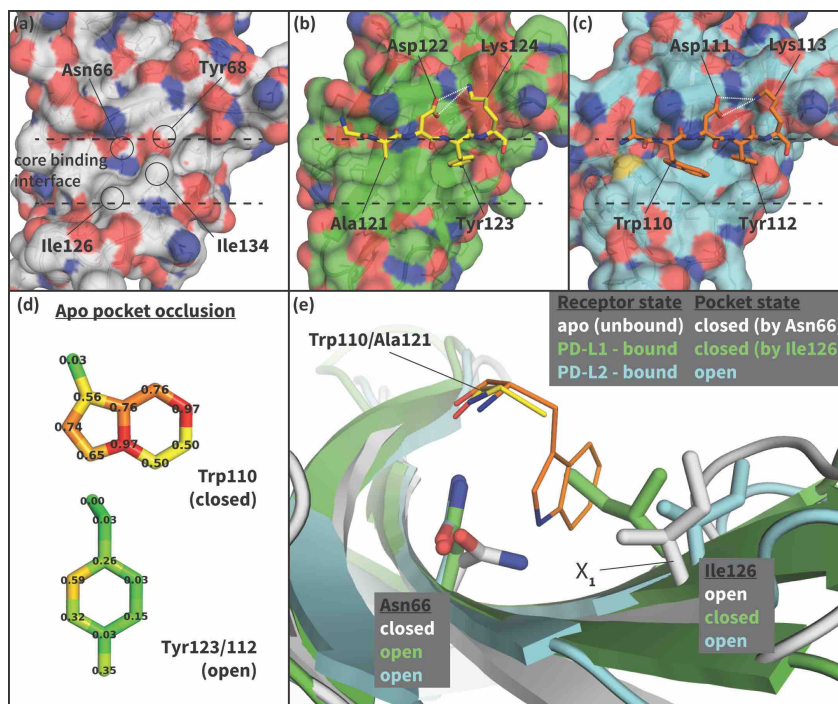


Figure 2.2. Flexibility of the PD-1 binding interface. (a) The apo PD-1 binding interface [19], showing a flat, polar, core binding interface. Surface residues that shape the core binding interface are labelled. (b) The core PD-1 (green) - PD-L1 (yellow) binding interface, showing a flat hydrophobic receptor surface [20]. White dashed lines indicate hydrogen bonds between PD-L1 side chains. (c) The core PD-1 (cyan) – PD-L2 (orange) binding interface, showing a large hydrophobic receptor cavity [21]. White dashed lines indicate hydrogen bonds between PD-L2 side chains. Note that the conserved anchor residue Tyr123/112 is present in both (b) and (c). (d) Fractional occlusion of each bound-like Trp110 and Tyr123/112 atom position in the NMR ensemble of apo PD-1. Numerical values at each atom position denote the fraction of NMR frames that overlap, or “occlude”, that position (see Chapter 2.4 Methods for full details of how fractional occlusion is calculated). Aside from the C_β, the Trp110 pocket is mostly occluded in the apo PD-1 ensemble, whereas the Tyr123/112 anchor pocket is largely open. (e) Overlay of apo, PD-L1 – bound, and PD-L2 – bound structures of PD-1 defining the “open” and “closed” states of PD-1 residues Asn66 and Ile126 in relation to the open and closed states of the Trp110 binding pocket.

To study the mechanism of PD-1 binding we use molecular dynamics simulations (MDs) to identify and quantify the effects of intermolecular interactions on the PD-1 binding interface. We first estimate ΔG_{BL}^{apo} for the free receptor and demonstrate that apo_{BL} states are exceedingly rare. We then estimate ΔG_{BL}^{EC} for PD-1 interacting with various peptide constructs that mimic distinct subsets of ligand interface motifs (Figure 2.3) and identify the critical features that trigger shifts in the PD-1 conformational ensemble towards the bound-like states. By quantifying the energetic contribution of each triggering contact in the EC_{NBL}, we rationalize how PD-1 uses flexibility to simultaneously achieve both promiscuity, i.e., binding to multiple ligands, and specificity. We show that a conserved set of three contacts in the PD-1 encounter complexes with PD-L1/2 progressively lowers the free energy of bound-like receptor states with respect to the non-bound-like state. These molecular triggers reshape the non-bound-like hydrophilic interface around Asn66 into a bound-like hydrophobic surface. A fourth contact that differs by a single atom stabilizes this surface into either a shallow patch that interacts with Ala121 in PD-L1, or a deep cavity that buries Trp110 in PD-L2.

We find that these triggers, which include the anchor Tyr123/112 in PD-L1/PD-L2 (Figure 2.2b,c,d) [29], are highly conserved across species [21] and drive quantitatively similar, kinetically efficient downhill binding pathways. The importance of these triggers is underscored by the PD-1 – targeting, anti-cancer antibody pembrolizumab, which evolved via a distinct evolutionary pathway yet, as we show, exploits some of the same triggering machinery as PD-1's natural ligands. Finally, we suggest how these induced-fit triggers could be used in rational, small-molecule drug discovery by studying the binding mode of a potent macrocyclic PD-1 inhibitor. Collectively, our findings demonstrate how nature exploits structural flexibility to achieve selective binding promiscuity in regulatory proteins.

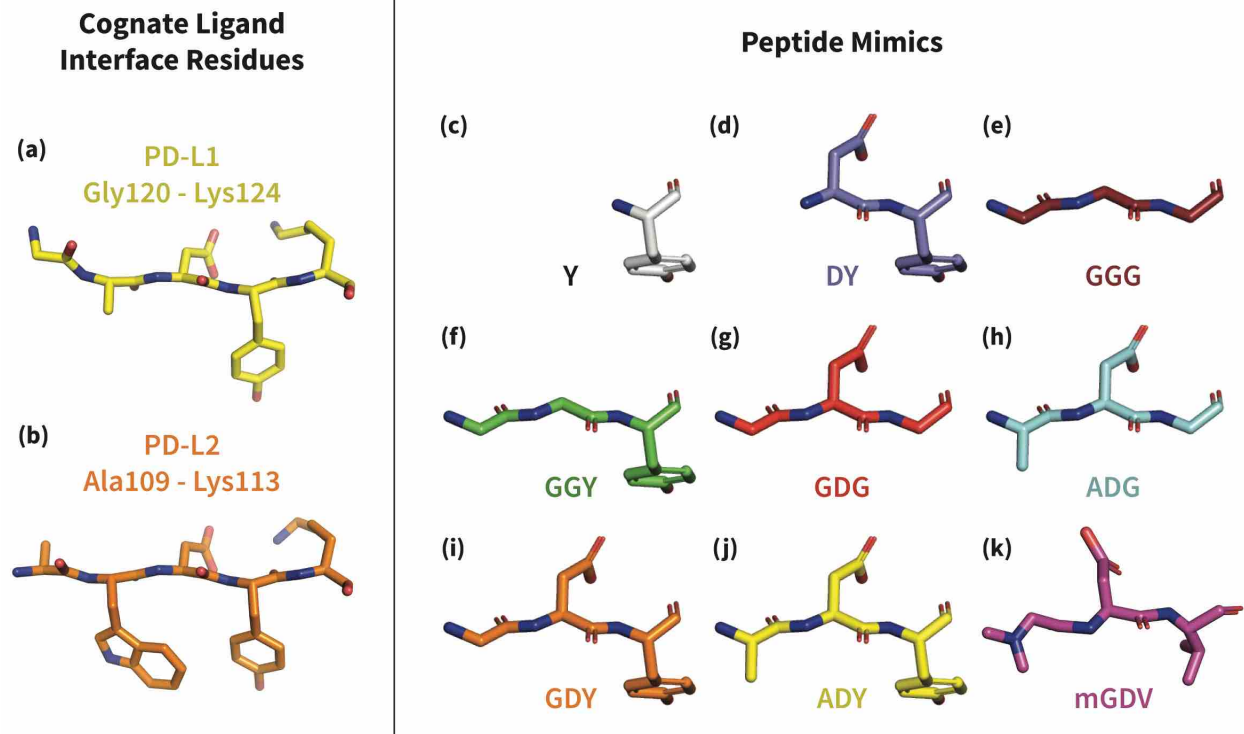


Figure 2.3. Structures of PD-L1/2 – mimicking peptides used to probe PD-1 interface dynamics. Left: core interface binding residues of (a) PD-L1 and (b) PD-L2 in their bound-like conformations. Right: peptides that were simulated in the presence of apo PD-1 in order to identify the triggers of induced fit interface deformations: (c) Y, (d) DY, (e) GGG, (f) GGY, (g) GDG, (h) ADG, (i) GDY, (j) ADY, and (k) mGDV.

2.2 RESULTS

2.2.1 Open and closed states of PD-1 Asn66 and Ile126 describe a hydrophilic or hydrophobic interface.

Analysis of aligned PD-1 structures (Figure 2.2) led us to classify the bound-like and non-bound-like conformational states using two binary order parameters defined by the 'open' or 'closed' states of Asn66 and Ile126. Namely, for a non-bound-like interface Asn66 is closed and Ile126 is open; for the PD-L1-specific bound-like state Asn66 is open and Ile126 is closed; and for the PD-L2-specific bound-like state both Asn66 and Ile126 are open (Figure 2.2e). In the PD-L1 – bound state, the interface exhibits a large hydrophobic patch that interacts with the side chain of ligand interface residue Ala121 (Figure 2.2b). In the PD-L2 – bound state, the interface exhibits a deep hydrophobic cavity that buries ligand residue Trp110 (Figure 2.2c). Neither this hydrophobic patch nor deep cavity is sampled in the apo PD-1 NMR ensemble, where, instead, the closed state of Asn66 blocks the Trp110 binding pocket by exposing its NH₂ group (Figure 2.2a,e, Figure 2.14), making a hydrophilic site. MDs of apo PD-1 confirm that Asn66 remains closed (Figure 2.4a), stabilized by a hydrogen bond with Lys78 that is also present in NMR structures (Figure 2.5a). These findings suggest that specific interactions between apo PD-1 and a nearby ligand might be required to open Asn66 and reshape the hydrophilic interface into its hydrophobic bound-like states.

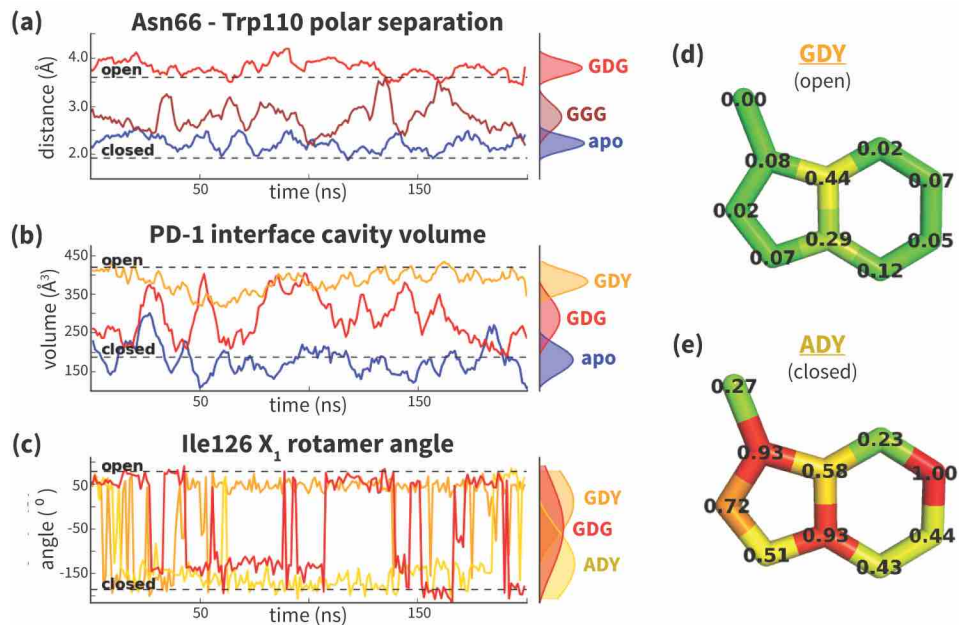


Figure 2.4. Dynamics of PD-1 binding interface in the presence of different ligands. (a) Rolling averages of distance between Trp110_NE1 (from bound PD-L2) and Asn66_ND2 from MDs of apo PD-1 (blue) alone and interacting with GGG (maroon) and GDG (red) peptides. Only GDG peptide sequesters Asn66 away from Trp110 binding pocket. (b) Rolling averages of PD-1 binding cavity volume from simulations of apo PD-1 alone (blue) and interacting with GDG (red) and GDY (orange) peptides shows that only GDY stabilizes an open cavity. (c) Ile126 X_1 rotamer angle from MDs of apo PD-1 interacting with GDG (red), GDY (orange), and ADY (yellow) peptides. Peptide ADY and GDY position Ile126 in the closed and open states, respectively (as in Figure 2.2e). Replicate trajectories for panels a, b, and c are shown in Figure 2.16. (d) Fractional occlusion of each bound-like Trp110 atom position in simulations of PD-1 interacting with the GDY peptide show an open Trp110 binding pocket. The fractional occlusion of a Trp110 atom position is defined as the percentage of simulation frames in which a PD-1 atom overlaps, or “occludes”, that position (see Chapter 2.4 Methods for full details of how fractional occlusion is calculated). (e) Fractional occlusion of each bound-like Trp110 atom position in simulations of PD-1 interacting with the ADY peptide show a closed Trp110 binding pocket.

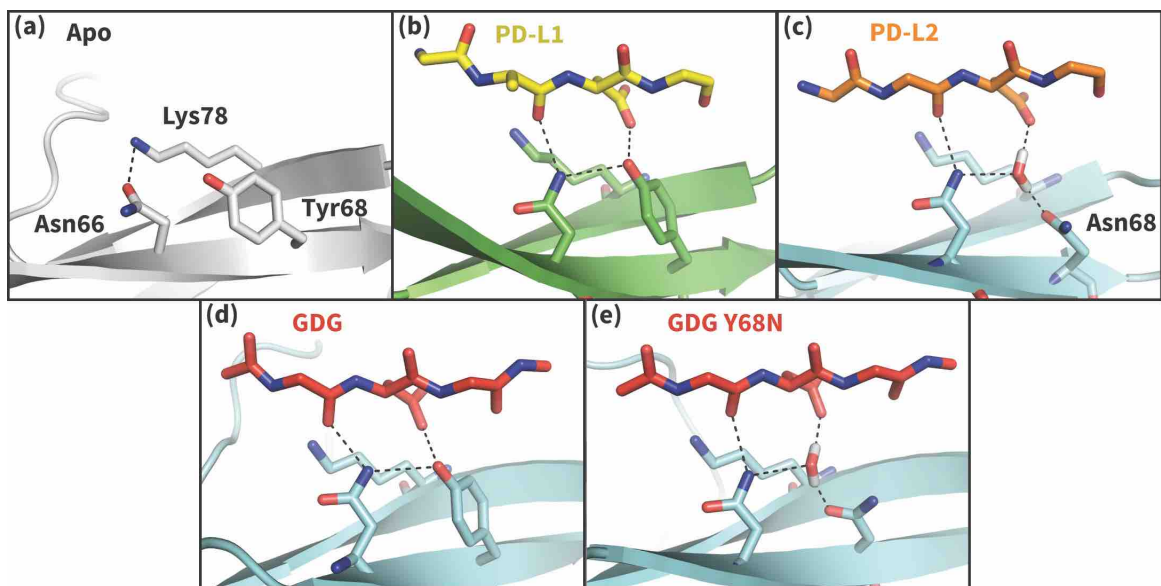


Figure 2.5. Hydrogen bond network of PD-1 Asn66 in different contexts. (a) NMR structure of the dominant apo, non-bound-like state of the human PD-1 interface [19]. Asn66 is in the closed state, forming a single hydrogen bond with Lys78. (b) Cocrystal structure of the human PD-1 – PD-L1 complex [20]. PD-1 bound-like interface shows Asn66 in the open state, forming two hydrogen bonds with the ligand Ala121 backbone and the neighboring Tyr68. For clarity only relevant ligand atoms are shown. (c) Cocrystal structure of the murine PD-1 – PD-L2 complex [21]. PD-1 bound-like interface shows Asn66 is in the open state, forming two hydrogen bonds with the ligand Trp110 backbone and a crystal water stabilized by neighboring residue Asn68. (d) Simulation snapshot of human PD-1 interacting with the GDG peptide, showing the same hydrogen bond network as in (b). (e) Simulation snapshot of human PD-1 Y68N mutant interacting with the GDG peptide, showing the same water-mediated hydrogen bond network as in (c).

2.2.2 Bound-like conformations of unbound Tyr123/112 in PD-L1/2 facilitates molecular recognition.

For both induced fit and conformational selection, the association of the apo receptor and ligand is driven mainly by diffusion [37, 38]. It has been shown that often protein-protein interactions stabilize the initial encounter complex through the burial of a bound-like anchor motif on the ligand [29], which allows subsequent, longer-timescale intermolecular interactions to take shape. Co-crystal structures, MDs and docking studies of PD-L1/2 suggest that the homologous interface residues Tyr123/112 (see Figure 2.2b,c) may serve as anchors. Specifically, MDs of apo PD-L1/2 show that Tyr123/112 remain within 0.5 Å heavy atom RMSD of their bound-like conformations 88 ± 16 % of the time. Furthermore, the Tyr123/112 binding pocket is unobstructed in the apo PD-1 NMR ensemble (Figure 2.2d), facilitating immediate burial of the side chain upon association. Docking exercises also point to the stabilizing role of the Tyr anchor. Namely, ClusPro [39] successfully re-docked the wild-type human PD-1 – PD-L1 co-crystal [20], but it failed for single-residue PD-L1 mutants Y123G and Y123A (Table 2.1). Collectively, these results suggest an anchor role for Tyr123/112 that facilitates molecular recognition between non-bound-like apo PD-1 and its ligands (as sketched in Figure 2.1).

Table 2.1. Anchor Tyr123 is key determinant of bound-like docked conformations. Backbone RMSD of top 10 ClusPro [39] predicted PD-L1 binding modes to the human PD-1 – PD-L1 cocrystal (PDB: 4ZQK). RMSDs shown for docked wild type human PD-L1 (WT) and for docked PD-L1 anchor mutants Y123G and Y123A.

ClusPro Model	Docked PD-L1 Backbone RMSD (Å) to 4ZQK PD-L1		
	WT	Y123G	Y123A
0	4.65	8.8	49.7
1	54.0	38.2	49.1
2	49.5	49.1	39.2
3	47.5	40.4	40.4
4	39.4	49.4	48.5
5	48.0	40.07	53.2
6	45.8	53.2	49.5
7	40.6	46.5	48.1
8	48.6	47.8	47.6
9	50.7	48.7	50.4

2.2.3 Conserved PD-L1/2 Asp122/111 form a specific intermolecular hydrogen bond network that opens PD-1 Asn66 and switches the receptor interface from hydrophilic to hydrophobic.

Co-crystal structures of bound PD-1 exhibit an open Asn66 that forms two hydrogen bonds: the first with the backbone oxygen of homologous PD-L1/2 Ala121/Trp110, and the second with either PD-1 Tyr68 (human PD-1 - PD-L1 complex) or a crystal water (murine PD-1 – PD-L2 complex) (Figure 2.5b,c). MDs of PD-1 in complex with a GGG peptide positioned to mimic the backbone of PD-L1/2 residues ADY123 and WDY112, respectively, show that Asn66 fluctuates back and forth between a bound-like open state, where it makes the aforementioned backbone hydrogen bond to the GGG peptide, and the non-bound-like closed state, where it is bonded to PD-1 Lys78 (Figure 2.4a). On the other hand, simulations with a GDG peptide show that the Asp122/111 mimic forms a hydrogen bond to the Tyr68 OH group, stabilizing a

Tyr68 rotamer that can simultaneously hydrogen bond to the NH2 of Asn66 (Figure 2.5d). Together, this Asn66 – Tyr68 hydrogen bond and the aforementioned Asn66 – backbone hydrogen bond stabilize the bound-like open state of Asn66 (Figure 2.4a).

The robust, four-membered hydrogen bond network between the Ala121/Trp110 backbone mimic, Asn66, Tyr68, and the Asp122/111 mimic that we observe in GDG MDs is fully consistent with all available structures and mutagenesis experiments. Namely, the hydrogen bonds rationalize the conservation of Asp122/111 in all known PD-L1/2 sequences and explain PD-L2 mutagenesis studies showing that the D111A mutation abolishes binding to PD-1 [21]. MDs of apo PD-L1/2 further support the importance of Asp122/111 interactions in the encounter complex by showing that this side chain remains within 0.4 Å RMSD of its bound-like conformation 82 ± 25 % of the time. The stabilization of the bound-like Asp122/111 side chain in simulation is achieved via hydrogen bonds with the neighboring Lys124/113, bonds which are also observed in bound cocrystal structures of PD-1 (Figure 2.2b,c). The importance of this stabilizing interaction is underscored by the fact that the K124S and K113A point mutations in PD-L1 and PD-L2, respectively, both abolish binding to PD-1 [21, 22].

PD-1 ligands open Asn66 by offering two novel hydrogen bonds (with the Ala121/Trp110 backbone and Tyr68) that out-compete the single Lys78 hydrogen bond that stabilizes the closed state. Interestingly, the one known PD-1 sequence that diverges at the Tyr68 position is murine PD-1, which has a Y68N mutation. The murine PD-1 – PD-L2 co-crystal shows that although the shorter Asn68 side chain cannot hydrogen bond directly to Asn66 or Asp111, it hydrogen bonds to a crystal water molecule that forms the same hydrogen bond network as Tyr68 (Figure 2.5c). MDs of a human Y68N PD-1 mutant and the GDG peptide suggest a functional equivalence of Asn68 to Tyr68: the Asn68 side chain spontaneously recruits a stable

water to the co-crystal position that then opens Asn66 via a specific hydrogen bond network analogous to that formed by Tyr68 (Figure 2.5e).

2.2.4 ADY/GDY ligand motifs stabilize distinct bound-like states for PD-L1/2.

While GDG MDs show an open Asn66 (Figure 2.4a) that exposes a hydrophobic surface, this surface remains flexible and fluctuates between a deep open cavity and closed shallow patch (Figure 2.4b). Contrary to the GGG MDs that exhibited open-closed fluctuations of Asn66 (Figure 2.4a), the pocket instability observed in GDG MDs is caused by open-closed fluctuations of PD-1 residue Ile126 (Figure 2.4c). In contrast, MDs show that the GDY peptide stabilizes the open states of both Asn66 and Ile126 and maintains the open hydrophobic interface cavity seen in the PD-L2 bound-like state of PD-1 (Figure 2.4b,c,d). Comparison of the GDG and GDY MDs reveal that the Tyr side chain serves as a ‘wedge’ that stabilizes the flexible loop surrounding Ile134 into a bound-like configuration that is observed in both the PD-L1 and PD-L2 co-crystal structures (Figure 2.6). In the presence of the GDY peptide, the bound-like Ile134 makes a hydrophobic contact with the long arm of Ile126, which pulls the latter residue out of the pocket and stabilizes its open state (Figure 2.15).

Although the PD-L1 interface exhibits the GDY scaffold, Ile126 is closed in the PD-L1-specific EC_{BL} state, suggesting that an additional ligand motif not contained in the GDY scaffold is responsible for closing the pocket. MDs with an ADY peptide that mimics Ala121 show that the extra C_β carbon of the Ala side chain out-competes Ile134 for the long arm of Ile126, stabilizing its closed state (Figure 2.4c,e). Interestingly, MDs with GDG and ADG peptides both show similarly unstable open-closed fluctuation of Ile126 (see Figure 2.7 below), which suggests that the effect of the Ala121 C_β carbon on Ile126 dynamics only emerges in the presence of the anchor Tyr123/112. Thus, in addition to facilitating molecular recognition,

stabilization of the Ile134 loop by the burial of Tyr123/112 is shown to enable ligand-specific induced fit responses by the PD-1 interface.

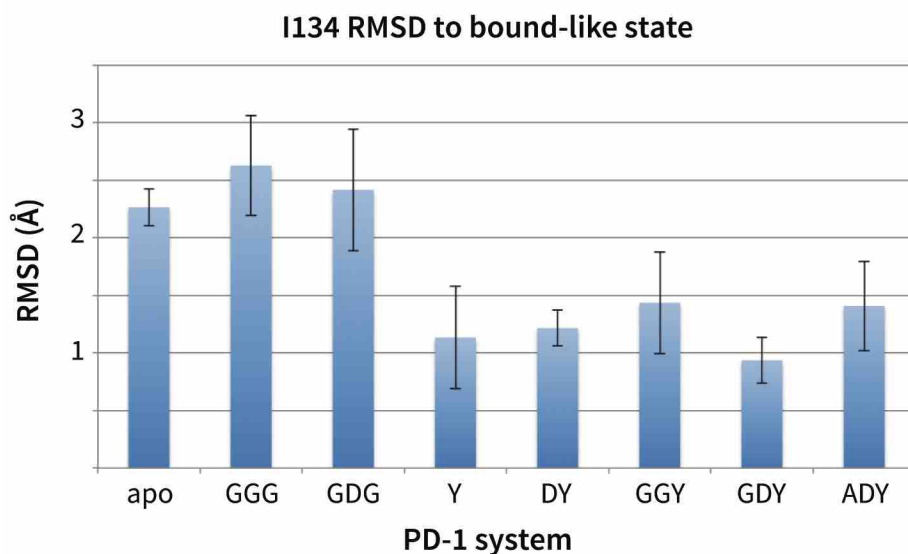


Figure 2.6. Stabilization of bound-like Ile134 by conserved tyrosine (Y) anchor. Average and standard deviation heavy atom RMSD of PD-1 Ile134 to the PD-L1/2 bound-like state (measured from human PD-1 – PD-L1 cocrystal, 4ZQK; Ile134 has $< 0.2 \text{ \AA}$ heavy atom RMSD between 4ZQK and the PD-L2 cocrystal 3BP5). Data is shown for three 200ns replicate simulations for each system, including apo human PD-1 and PD-1 interacting with various peptides.

Although the PD-L1 interface exhibits the GDY scaffold, Ile126 is closed in the PD-L1-specific EC_{BL} state, suggesting that an additional ligand motif not contained in the GDY scaffold is responsible for closing the pocket. MDs with an ADY peptide that mimics Ala121 show that the extra C_{β} carbon of the Ala side chain out-competes Ile134 for the long arm of Ile126, stabilizing its closed state (Figure 2.4c,e). Interestingly, MDs with GDG and ADG peptides both show similarly unstable open-closed fluctuation of Ile126 (see Figure 2.7 below), which suggests that the effect of the Ala121 C_{β} carbon on Ile126 dynamics only emerges in the presence of the anchor Tyr123/112. Thus, in addition to facilitating molecular recognition,

stabilization of the Ile134 loop by the burial of Tyr123/112 is shown to enable ligand-specific induced fit responses by the PD-1 interface.

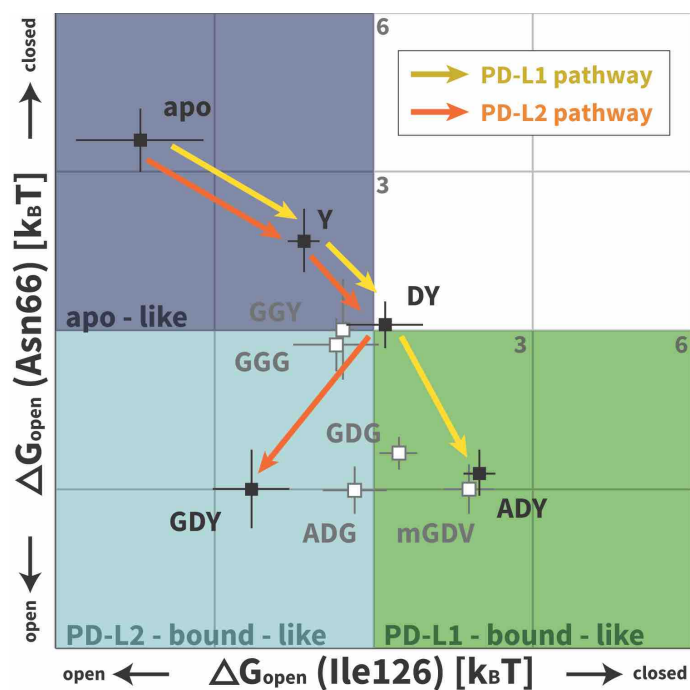


Figure 2.7. Downhill binding pathways of PD-1 triggers of induced fit for each cognate ligand. Points on the plot represent average and standard deviation equilibrium free energy differences (from three replicate simulations) between the open and closed states of receptor residues Asn66 and Ile126 for apo PD-1 and PD-1 interacting with nine distinct ligand-mimicking peptides. The corresponding numerical values can be found in Table 2.2. Yellow and orange lines represent the ligand-specific induced fit binding pathways from the apo receptor ensemble to the PD-L1 and PD-L2 bound-like ensembles, respectively.

Table 2.2. Free energy difference between the non-bound-like and bound-like states of PD-1 interface residues Asn66 and Ile126 in various systems. Listed values show the average and standard deviation of ΔG_{BL} (from three replicate simulations) for Asn66 and Ile126 in the different PD-1 systems. Since the bound-like state of Ile126 is closed when PD-L1 – bound and open when PD-L2 - bound, the ΔG_{BL} values for this residue take opposite signs. The trivial relationship between ΔG_{BL} and ΔG_{open} are indicated for each column. Values shown are in units of $k_B T$, with $T = 300$ K.

	PD-L1 / PD-L2	PD-L1	PD-L2
PD-1 Simulation	$\Delta G_{BL}(\text{Asn66})$ $\Delta G_{open}(\text{Asn66})$ ($k_B T$)	$\Delta G_{BL}(\text{Ile126})$ $-\Delta G_{open}(\text{Ile126})$ ($k_B T$)	$\Delta G_{BL}(\text{Ile126})$ $\Delta G_{open}(\text{Ile126})$ ($k_B T$)
apo (ΔG_{BL}^{apo})	3.6 ± 0.60	4.4 ± 1.2	-4.4 ± 1.2
Y (ΔG_{BL}^Y)	1.7 ± 0.61	1.3 ± 0.3	-1.3 ± 0.3
DY (ΔG_{BL}^{DY})	0.11 ± 0.44	-0.21 ± 0.74	0.21 ± 0.74
GGG (ΔG_{BL}^{GGG})	-0.28 ± 0.52	0.7 ± 0.8	-0.7 ± 0.8
GGY (ΔG_{BL}^{GGY})	0.02 ± 0.17	0.58 ± 0.22	-0.58 ± 0.22
GDG (ΔG_{BL}^{GDG})	-2.3 ± 0.32	-0.46 ± 0.36	0.46 ± 0.36
ADG (ΔG_{BL}^{ADG})	-3.0 ± 0.45	0.35 ± 0.60	-0.35 ± 0.60
GDY (ΔG_{BL}^{GDY})	-3.0 ± 0.7	2.3 ± 0.72	-2.3 ± 0.72
ADY (ΔG_{BL}^{ADY})	-2.7 ± 0.44	-2.0 ± 0.36	2.0 ± 0.36
mGDV (ΔG_{BL}^{mGDV})	-3.0 ± 0.47	-1.8 ± 0.48	1.8 ± 0.48

2.2.5 PD-L1/2 triggers ADY/GDY produce energetically downhill induced fit binding pathways.

We applied Maxwell-Boltzmann statistics to our peptide simulations (see Chapter 2.4 Methods) to quantify the role played by each trigger in the structural transitions at the PD-1 interface. We evaluate ΔG_{open} , i.e., the free energy differences between the open and closed states of Asn66/Ile126 for PD-1 in isolation and PD-1 interacting with nine different peptides representing distinct PD-L1/2 interface motifs (Figure 2.3, Table 2.2; note that ΔG_{open} and ΔG_{BL} are trivially related). These ΔG_{open} values are plotted in Figure 2.7. Remarkably, the ADY and GDY motifs respectively shift the ratio of our predefined bound-like to non-bound-like states from 1 : 44 ± 24 (based on $\Delta G_{open}^{apo}(Asn66)$) to 7.4 ± 2.8 : 1 for the PD-L1 bound-like state (based on $\Delta G_{open}^{ADY}(Ile126)$) and 12 ± 9.6 : 1 for the PD-L2 bound-like state (based on $\Delta G_{open}^{GDY}(Ile126)$). More importantly, we show that each triggering contact monotonically lowers the relative free energy of ligand-specific bound-like states starting from no contacts (apo), to the first, conserved contact with the anchor (Y), to the second, conserved contact with Asp122/111 (DY), to the final, unconserved contact with the backbone O of A/G in the complete triggering motifs (ADY/GDY) (Figure 2.7). The fact that these downhill binding pathways do not encounter energy barriers strongly suggests that the PD-1 binding mechanism is primarily one of induced fit (see Figure 2.1).

In the apo simulation Asn66 is closed ($\Delta G_{open}^{apo}(Asn66) \approx 3.6 k_B T$), repelling Ile126 into an open conformation ($\Delta G_{open}^{apo}(Ile126) \approx -4.4 k_B T$). Docking of the Tyr anchor (Y) and formation of the encounter complex destabilizes the non-bound-like apo PD-1 interface, causing increased open-closed fluctuations in both Asn66 and Ile126. The subsequent docking of Asp122/111 (DY) allows Tyr68 to compete with Lys78 to form one hydrogen bond with Asn66, causing it to swap back and forth between open and closed ($\Delta G_{open}^{DY}(Asn66) \approx 0$). Fluctuations of Asn66 correlate with simultaneous fluctuations of Ile126 ($\Delta G_{open}^{DY}(Ile126) \approx 0$). Adding the adjacent Ala121/Trp110 backbone from PD-L1/2 (ADY/GDY)

provides the second hydrogen bond for the NH₂ of Asn66 that fully stabilizes its open state ($\Delta G_{open}^{GDY/ADY}(Asn66) \approx -3.0 k_B T$). With Asn66 open, the C β atom of Ala121 modulates Ile126 dynamics. When present (ADY), the C β hydrophobically recruits Ile126 into the closed pocket state ($\Delta G_{open}^{ADY}(Ile126) \approx 2.0 k_B T$). Without C β (GDY), Ile126 remains open ($\Delta G_{open}^{GDY}(Ile126) \approx -2.3 k_B T$).

Our ΔG_{open} calculations also quantify the critical role of the anchor residue Tyr123/112 in ensuring the ligand specificity of PD-1 interface deformations. This is demonstrated by the fact that GDY and ADY peptides impose clear differential influence on the dominant rotamer state of Ile126, while for both GDG and ADG, Ile126 fluctuates about evenly between the open and closed state ($\Delta G_{open}^{GDG/ADG}(Ile126) \approx 0$) (Figure 2.7).

2.2.6 Encounter complex simulations suggest chronology of induced fit triggering interactions.

We ran MDs of the PD-L1/2 encounter complexes starting from docked poses of apo PD-1 and the interacting domains of PD-L1/2 that anchored Tyr123/Y112 (see Chapter 2.4 Methods). Encounter complex MDs recapitulated the triggering mechanisms we identified in our peptide simulations and their resulting PD-1 interface transitions from the EC_{NBL} to the ligand-specific EC_{BL} states. The chronology for these interactions (Table 2.3) is the same for both ligands. Consistently, the first interaction to take place after docking the conserved anchor is the formation of the hydrogen bond between receptor residue Tyr68 and ligand residue Asp122/111. This is followed by stabilization of Asn66 in the open pocket state via hydrogen bonds with neighboring Tyr68 and the ligand Ala121/Trp110 backbone. The Ala121/Trp110 side chains then proceed to stabilize a closed/open hydrophobic pocket. Note that the Trp in the WDY motif of PD-L2 readily fills the hydrophobic pocket as the XDY motif latches and opens Asn66 (Figure 2.8). Consistent with a downhill free energy induced fit mechanism, the realization of these four contacts takes less than 10 ns total. On a longer timescale, encounter complex simulations demonstrate the formation

of secondary hydrogen bonds at the interface periphery that are also observed in co-crystal structures of human and murine PD-1. These secondary hydrogen bonds, including the bond from PD-1 Lys78 to PD-L1/2 Phe19/21 and from Gln75 to Arg125/Tyr114 (Figure 2.9), were consistently observed to form approximately 10 nanoseconds after the aforementioned Asn66 and Tyr68 hydrogen bonds (Table 2.3), suggesting that EC_{BL} contacts shaped by the triggers of induced fit are enough to drive the subsequent transition to the HAC.

Table 2.3. Chronology of the formation of intermolecular interactions between PD-1 and PD-L1/2 in encounter complex simulations. Listed values show the average and standard deviation time to formation (from three replicate simulations) of various inter- and intra-molecular hydrogen bonds following the burial of the ligand anchor and formation of the key Tyr68–Asp122/111 hydrogen bond.

Hydrogen Bond	Δt (ns) after Tyr68 – Asp122/111 hydrogen bond formation	
	PD-1 - PD-L1 Encounter Complex	PD-1 - PD-L2 Encounter Complex
Asn66 – Ala121/Trp110	6.3 ± 2.9	6.7 ± 7.2
Asn66 – Tyr68	5.0 ± 1.7	8.3 ± 7.5
Gln75 – Arg125/Tyr114	15 ± 7.8	17 ± 11
Lys79 – Phe19/21	13 ± 15	15 ± 20

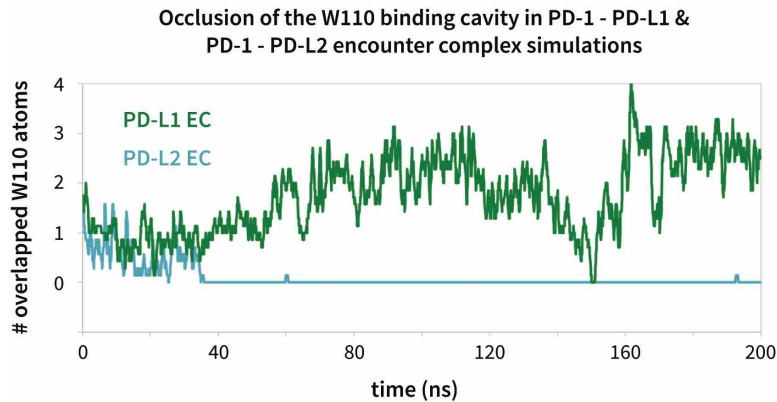


Figure 2.8. Modulation of the PD-1 interface binding cavity in encounter complex simulations with PD-L1 and PD-L2. Plot shows the (rolling average) number of atoms in the bound-like Trp110 side chain reference that are occluded by the PD-1 interface throughout encounter complex simulations with PD-L1/2 (see Chapter 2.4 Methods for full details of how occlusion is calculated). Both encounter complexes begin with a closed Trp110 pocket, as this is the dominant state of apo PD-1. The PD-L2 trigger then stabilizes the hydrophobic cavity (no overlap), while the PD-L1 trigger stabilizes the hydrophobic patch (significant overlap).

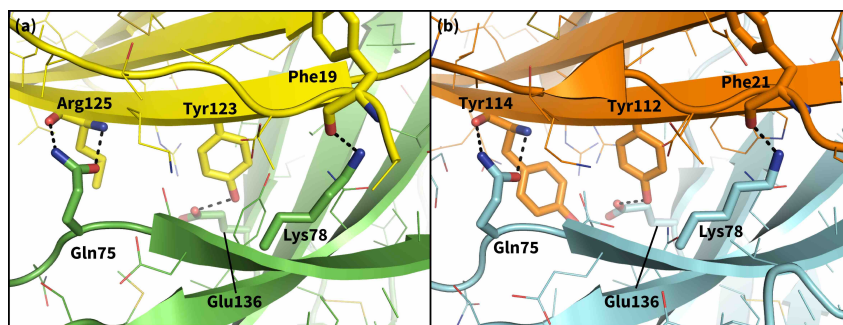


Figure 2.9. Secondary, non-triggering contacts in PD-1 encounter complexes. Specific hydrogen bonds observed in the PD-1 – PD-L1 (a) [20] and PD-1 – PD-L2 (b) [21] cocrystal structures. In simulation these contacts form approximately 10 ns after triggering interactions and their resulting induced fit deformations of the receptor (Table 3). Note also that the conserved Tyr123/112 anchor forms identical hydrogen bonds with Glu136 in the PD-L1 – and PD-L2 – bound states.

2.2.7 PD-1 – targeting antibody validates the critical role of Asn66 and suggests an anchor-independent binding mechanism with closed Ile126 and Ile134.

Recently, two FDA-approved PD-1 – targeting antibodies have emerged as part of a new generation of anti-cancer immune checkpoint inhibitors. Published crystal structures of one of these antibodies, pembrolizumab, bound to extracellular PD-1 show a hydrophobic receptor binding surface that overlaps that which binds PD-L1/2 (Figure 2.10b) [40-42]. Comparison of the pembrolizumab – PD-1 interface to the PD-L1 – PD-1 interface using the FastContact server [43] highlights several differences in the main contacts that characterize the two binding modes (Figure 2.10a, Tables 2.4, 2.5). Remarkably, the pembrolizumab-bound crystal structures reveal that the antibody stabilizes the same open state of Asn66 as PD-L1/2 using an analogous hydrogen bond network (Figure 2.10c). The fact that this antibody, designed via a distinct evolutionary pathway, shares PD-L1/2's mechanism for opening Asn66 and revealing a hydrophobic binding surface (Figure 2.2a,b,c, Figure 2.10b) underscores the role of this specific interaction in PD-1 interface remodeling.

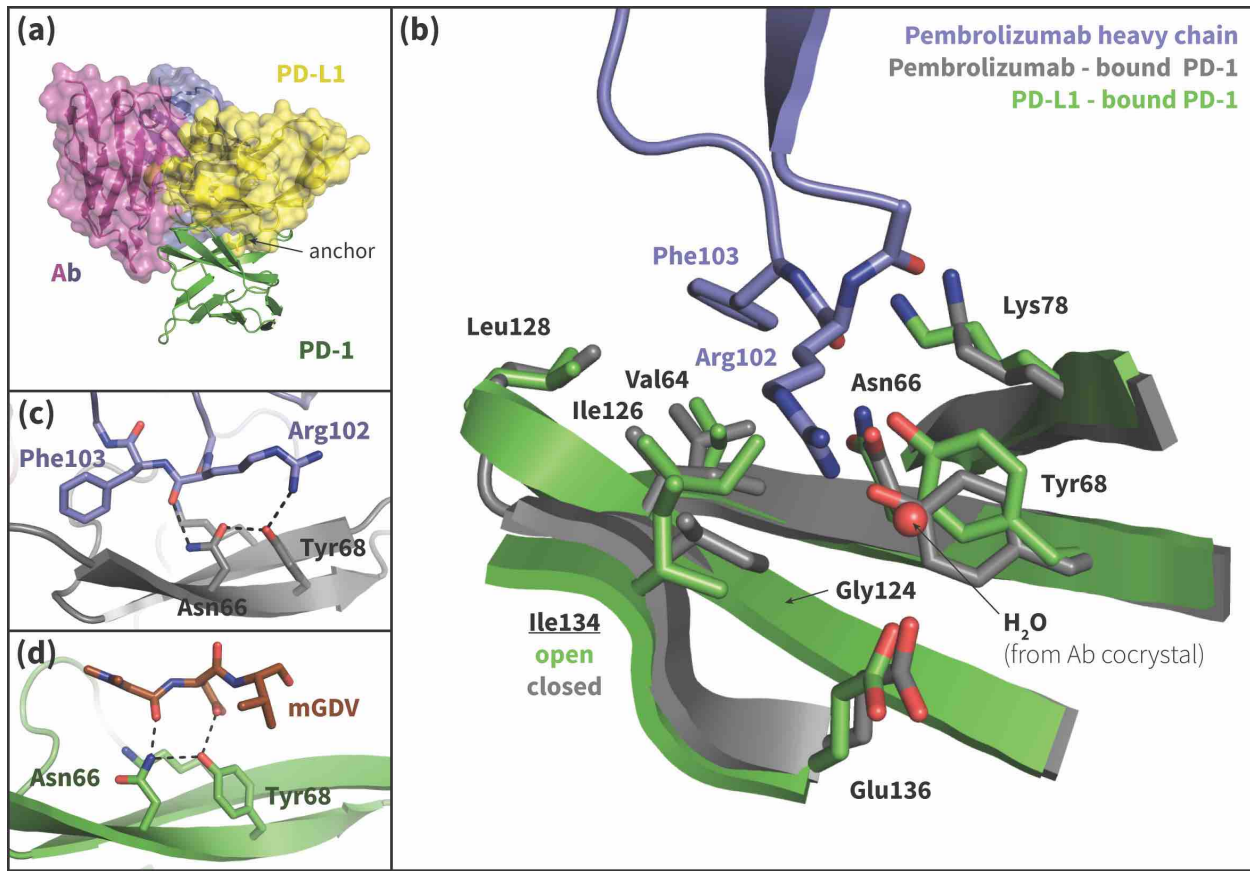


Figure 2.10. Pembrolizumab – bound PD-1 interface resembles PD-L1 – bound interface with a closed Ile134. (a) Alignment of crystal structures of the pembrolizumab antibody (Ab) [40] and PD-L1 [20] binding modes, showing distinct but partially overlapping binding interfaces on PD-1. The light chain of the Ab is shown in magenta and the heavy chain is shown in purple. (b) Detailed comparison of the aligned Ab – bound (grey) and PD-L1 – bound (green) PD-1 interfaces. Most receptor interface residues exhibit near-identical conformations, except Ile134 which is open when bound to PD-L1 but closed when bound to pembrolizumab. Heavy chain Ab interface residues are shown in purple. (c) Detail of the Ab – PD-1 interface, highlighting the hydrogen bond (hydrogen bond) network that stabilizes the open state of Asn66. This hydrogen bond network is functionally analogous to those observed in the PD-L1 and PD-L2 – bound cocrystals (Figure 2.5), although the OD1 and ND2 atoms of Asn66 are flipped. (d) Simulation snapshot of human PD-1 interacting with the mGDV motif from Bristol-Myers Squibb macrocyclic PD-1 inhibitor, highlighting the canonical hydrogen bond network that opens Asn66.

Table 2.4. Top 5 PD-1 residues contributing to electrostatic energy when binding to PD-L1 and pembrolizumab. Binding energies were calculated using the FastContact web server [43] and cocrystal structures of PD-1 bound to PD-L1 [20] and pembrolizumab [40].

PD-L1 – bound		Pembrolizumab – bound	
Residue	Energy (kcal/mol)	Residue	Energy (kcal/mol)
Glu136 ¹	-11.531	Asp85 ³	-8.367
Asp77	-5.073	Ser87	-3.629
Lys78 ²	-4.266	Asp77	-2.417
Gln75	-4.027	Tyr68	-2.156
Glu84	-3.119	Glu136	-2.096

¹ The E136A mutation abolishes binding of PD-1 to PD-L1 and greatly reduces binding to PD-L2 [21].

² The K78A mutation abolishes binding of PD-1 to PD-L1 and greatly reduces binding to PD-L2 [21].

³ The D85G mutation abolishes binding of PD-1 to pembrolizumab [42].

Table 2.5. Top 5 PD-1 residues contributing to desolvation energy when binding to PD-L1 and pembrolizumab. Binding energies were calculated using the FastContact web server [43] and cocrystal structures of PD-1 bound to PD-L1 [20] and pembrolizumab [40].

PD-L1 – bound		Pembrolizumab – bound	
Residue	Energy (kcal/mol)	Residue	Energy (kcal/mol)
Ile126 ¹	-1.853	Leu128 ²	-2.886
Leu128 ²	-1.673	Pro89	-2.486
Ile134 ³	-1.361	Val64	-1.721
Val64	-0.463	Pro130	-1.586
Ala132	-0.37	Pro83	-1.131

¹ The I126A mutation greatly reduces binding of PD-1 to both PD-L1 and PD-L2 [21].

² The L128A mutation abolishes binding of PD-1 to PD-L1 and partially reduces binding to PD-L2 [21].

³ The I134A mutation abolishes binding of PD-1 to PD-L1 and greatly reduces binding to PD-L2 [21].

Although pembrolizumab's interaction with Asn66 mimics the native-like contacts of PD-L1/2, the antibody-bound receptor exhibits a novel configuration of Ile134, with both Ile126 and Ile134 in inward-flipped, 'closed' states (Figure 2.10b). The result is a large hydrophobic surface where, like in the PD-L1 – bound state, the closed Ile126 occludes the Trp110 binding pocket, but where, unlike the PD-L1/2 – bound states, a closed Ile134 partially fills the Tyr/123/112 anchor cavity. In fact, pembrolizumab has no anchor analog. Instead, the Arg102 side chain extends along the PD-1 interface such that the C_z carbon overlaps the C_γ position of Tyr123/112 (Figure 2.18), and the NH1/2 groups hydrogen bond to a crystal water above the receptor interface (Figure 2.10b). In this configuration, the hydrophobic carbon chain of Arg102 forms a 'cap' above the closed Ile126 and Ile134, desolvating their hydrophobic interactions with each other and the neighboring Gly124 and stabilizing a flat hydrophobic surface (Figure 2.10b).

A similar closed conformation of Ile134 is observed in our MDs of PD-1 interacting with the GDG peptide (Figure 2.15). This is unsurprising: like pembrolizumab, the GDG peptide has the necessary machinery to trigger the opening of Asn66, but lacks an anchor 'wedge' that prevents the resulting inward collapse of Ile134. Results of the GDG MDs thus rationalize the pembrolizumab binding mode and suggest an anchor-independent induced fit PD-1 binding pathway: one in which the antibody opens Asn66 using the canonical hydrogen bond network and stabilizes the resulting flat hydrophobic interface by 'capping' the closed states of Ile126/134 with the carbon chain of Arg102.

2.2.8 Can molecular triggers be exploited to drug PD-1?

Although two PD-1 targeting antibodies already exist on the market, there are no small-molecule PD-1 inhibitors in clinical trial, despite the enormous interest in this blockbuster immunotherapy target [30-32]. Given that ligand binding sites tend to be concave [33, 34] and largely hydrophobic [35], the undruggability of PD-1 might be due to the closed Asn66 and the resulting flat polar interface in the apo

form (Figure 2.2a). However, the highly specific hydrogen bond network presented by PD-L1/2 and pembrolizumab strongly suggests a path to open Asn66 and transform the hard to drug hydrophilic patch into a hydrophobic one. Interestingly, Bristol-Myers-Squibb recently patented a 1.03 nM macrocyclic inhibitor of the PD-1 – PD-L1 interaction [44]. Although no mechanism of action has been described, the macrocycle includes a peptidic mGDV motif that is structurally similar to the aforementioned ADY induced fit trigger, with an N-methylated Gly and an Asp side chain that resemble PD-L1's Ala121 and Asp122, respectively (Figure 2.18, Figure 2.19). This alignment puts the mGDV motif's short Val side chain at the position of the much longer Tyr123 anchor, where it aligns with the C_Δ side chain carbon of pembrolizumab residue Arg112 (Figure 2.18).

Given the resemblance of the mGDV motif to the interface residues of both PD-L1 and pembrolizumab, we used our MDs method to evaluate whether this motif was capable of remodeling the apo, non-bound-like PD-1 interface into a bound-like state. We observed that mGDV opened Asn66 using a native-like hydrogen bond network analogous to those seen in previous simulations (Figure 2.5, 2.7, 2.10d). However, Ile126 and Ile134 dynamics mirrored those seen in the pembrolizumab cocrystal, with both sidechains favoring inward-flipped 'closed' configurations (Figure 2.11). Simulation trajectories showed that the short Val side chain of the mGDV motif, unlike the cognate Tyr123/112 anchors, did not penetrate deep enough into the PD-1 interface to be a 'wedge' stabilizing an open Ile134. Instead, like the carbon chain of pembrolizumab residue Arg102, the Val 'capped' stable hydrophobic interactions between a closed Ile134, a closed Ile126, and the neighboring Gly124.

PD-1 Ile126/134 χ_1 rotamer distributions

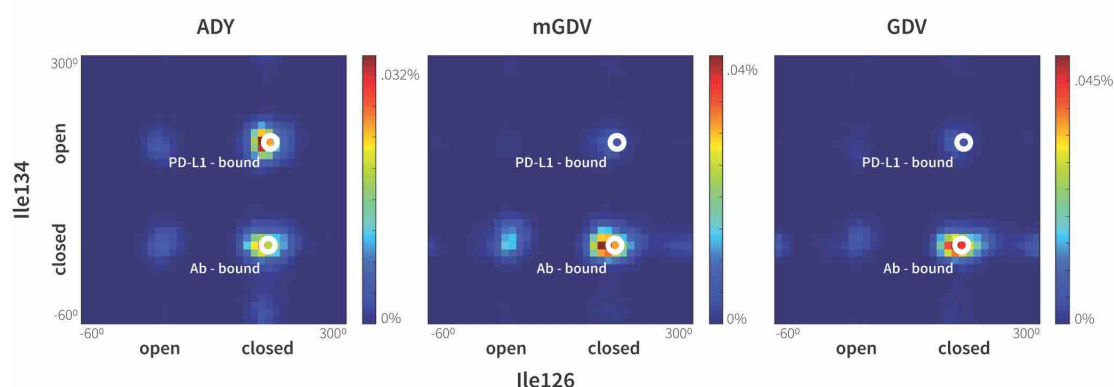


Figure 2.11. Macrocytic mGDV motif induces structural changes in the PD-1 interface towards the pembrolizumab – bound state. Heat maps show the distributions of PD-1's Ile126 and Ile134 χ_1 rotamer angles in MDs of the receptor interacting with the ADY PD-L1 trigger (left), the BMS macrocycle mGDV motif (center), and the GDV peptide. Data for each ligand was gathered from three distinct 200ns simulations. White dots on the plots indicate the rotamer angles of the same two residues in the pembrolizumab (Ab) – bound [40] and PD-L1 – bound [20] cocrystal structures.

Our GDG, ADG, GDY, and ADY simulations demonstrated that precise regulation of the closed/open states Ile126 via the Ala121 C_β is realized only when the Tyr123/112 anchor is buried (Figure 2.7). Thus, given that mGDV lacks an anchor, a natural question to ask is whether a Ile126 would be opened by a GDV peptide without the N-methyl group. Interestingly, MDs of PD-1 interacting with a GDV peptide revealed identical Ile126 and Ile134 dynamics to mGDV simulations (Figure 2.11), indicating that the N-methyl group was not recruiting Ile126 into the closed state in the style of Ala121 C_β . These results help to further illuminate the role of the conserved anchor Tyr123/112, which in its absence does not wedge Ile134 into the open state, disabling the capability of PD-1 to stabilize an open Ile126 and form a hydrophobic cavity at that site.

Compared to GDG simulations in which Ile126 fluctuated between open and closed (Figure 2.4b,c), in GDV simulations it remained closed, suggesting a stabilizing role for the Val side chain. The overlap of (m)GDV's Val with the carbon chain of pembrolizumab's Arg102 (Figure 2.18) and the similarity between the (m)GDV-induced PD-1 interface and the pembrolizumab – bound interface supports the 'capping' role of Arg102 in stabilizing the flat hydrophobic surface of PD-1. This mechanism is also consistent with models of macrocycle conformations generated by Balloon [45] docked to PD-1, which readily identify poses that align the mGDV motif to corresponding PD-L1 and pembrolizumab interface residues (Figure 2.18, Figure 2.19), rationalizing the potency and specificity of the compound.

2.3 DISCUSSION

2.3.1 Induced fit motif XDY shared by PD-1 ligands modulates the flexible PD-1 binding interface from hydrophilic to hydrophobic.

Our studies show that apo PD-1 does not sample bound-like hydrophobic interface conformations, but instead presents a non-bound-like hydrophilic patch around Asn66 at the core of its binding interface (Figure 2.2). By mapping the effect of specific ligand contacts on the apo PD-1 interface, we identify a highly-conserved subset of PD-L1/2 motifs responsible for coordinating Asn66 and triggering the transition from the hydrophilic to hydrophobic interface. Namely, Asp122/111 and the backbone O of PD-L1/2 Ala121/Trp110 form a robust, four-membered hydrogen bond network with Tyr68 and Asn66 that neutralizes the latter residue into a bound-like open state. Simultaneously, the conserved anchor Tyr123/112 stabilizes Ile134 into a bound-like state that, with Asn66 open, creates a hydrophobic surface that fluctuates between a patch and a cavity modulated by Ile126. These three linear ligand motifs (XDY), shared by both PD-L1/2, comprise the molecular key that unlocks the promiscuity of PD-1 by revealing a flexible hydrophobic binding surface (Figure 2.4).

2.3.2 A single carbon atom difference can shift the hydrophobic PD-1 binding surface from a stable patch to a stable cavity.

With XDY triggering the transition to the flexible hydrophobic surface, specificity towards the two PD-1 ligands is actualized by the formation of the hydrophobic patch when binding PD-L1 vs. the formation of hydrophobic cavity when binding PD-L2. These two states can be distinguished by the conformation of Ile126 (Figure 2.2e). For PD-L1, we show that the ADY motif is sufficient to stabilize the hydrophobic patch (Figure 2.4c). Specifically, the Ala121 C β atom, which does not overlap with PD-1 apo NMR structures (Figure 2.2d), recruits Ile126 into the closed (patch) state. On the other hand, in the absence of C β , the GDY trigger stabilizes the open state of Ile126, producing a large hydrophobic interface cavity consistent with the pocket that buries PD-L2 Trp110. Note that the Trp in the WDY motif of PD-L2 readily fills the hydrophobic pocket as the XDY motif latches and opens Asn66 (Figure 2.8).

2.3.3 Bound-like XDY residues and molecular recognition.

The pre-arrangement of PD-L1/2 motifs XDY in bound-like conformations in the absence of the receptor is important for efficient ligand recognition and binding. Docking studies and peptide MDs highlight a critical role for the conserved Tyr123/112 anchor both in both molecular recognition and in modulating Ile134 during induced fit, both of which require the Tyr side chain to maintain a stable bound-like rotamer. Furthermore, simulations demonstrate that peptides such as GDG, mGDV, and GDV, which either lack or have a modified anchor analogue, cannot stabilize an open state of Ile126, highlighting an allosteric role for Tyr123/112 in splitting the PD-1 induced fit binding pathway.

Several anchors substitutes were tested in simulation starting in bound-like configurations similar to the cognate Tyr112/123. These MDs produced three broad types of PD-1 interface dynamics (Figure 2.17): (1)

aromatic substitutions XDF and XDW stabilized either an open (X=G) or closed (X=A) pocket like the cognate XDY motif. (2) Polar substitutions XDH, XDR, and XDK were not accommodated in the hydrophobic anchor pocket and their side chains laid along the receptor surface, consistent with pembrolizumab's bound Arg102 (Figure 2.10b), producing a closed pocket like that of (m)GDV. (3) XDG or XDA resulted in open-closed fluctuations of both Ile134 and Ile126 (Figure 2.4b,c). These observations suggest that certain anchor mutations are tolerated by PD-1 and are consistent with mutagenesis studies showing that the Y112A PD-L2 point mutation slightly reduces, but does not abolish, binding to PD-1 [21]. However, the observed conservation of Tyr123/112 in mammalian species [21] might suggest specific kinetic constraints on ligand recognition arising from hydrophobic contacts with Ile134 and the hydrogen bond with Glu136 (Figure 2.9), which are not shared by other sidechains.

In addition to the anchor residue, our peptide MDs also suggest an essential role for the conserved Asp122/111 in erecting a stable hydrogen bond network that opens PD-1 Asn66, which can only be achieved by a bound-like Asp side chain. The primacy of these intermolecular interactions to PD-1 binding is reinforced by our MDs of apo PD-L1/2, which reveal that Tyr123/112 and Asp122/111 all remain in bound-like conformations in the absence of the receptor, primed to interact immediately upon interface association. Equally important is the fact that apo PD-1 structures all accommodate (i.e. do not block) any of contacts of the XDY scaffold, ensuring a rapid recognition process that facilitates subsequent induced fit transitions.

2.3.4 Downhill binding pathways strongly suggest an induced fit binding mechanism.

Our MDs demonstrate that the set of consecutive intermolecular interactions triggered by ADY and GDY peptides lead to energetically downhill binding pathways with no opposing energy barriers. These pathways strongly suggest that PD-1 occurs mostly by induced fit (Figure 2.1). Specifically, simulations and

estimated ΔG_{open} values show that apo_{BL} states of PD-1 are rare, which undermines a conformational selection mechanism. On the other hand, ligand-specific triggers are shown to efficiently shift the PD-1 interface conformational ensemble from a non-bound-like : bound-like ratio of roughly 44 : 1 (in the apo ensemble) to roughly 1 : 7 (in the encounter complex ensemble) (Figure 2.7). Unconstrained MDs of PD-L1/2 encounter complexes show that the geometry and chronology of triggering contacts is highly optimized, driving the transition from the non-bound-like to the bound-like states in less than 10 ns. This time scale promotes rapid recognition and ensures fast activation of this important T-cell checkpoint.

2.3.5 Two step binding pathway of PD-1 reveals a simple mechanism for selective promiscuity.

Although regulatory proteins are promiscuous in that they bind multiple targets, they must also be specific so as to limit binding to just those targets. Our analysis of the binding mechanism to PD-1 reveals how these two seemingly contradictory requirements can be simultaneously achieved. Here, we show that apo PD-1 samples an ensemble of non-bound-like conformations that present an obstructive Asn66 on its interface, which likely prevents non-specific binding. The apo PDL1/2 interfaces feature a conserved, bound-like, XDY binding motif that holds the key to opening Asn66 and forming a flexible hydrophobic surface, which completes the first binding step. In the second step, the ligands then attune the flexible interface via specificity-determinant contacts (X=A for PD-L1, X=W for PD-L2) that modulate Ile126, splitting the binding pathway and stabilizing either a hydrophobic patch or a binding pocket (Figures 2.2, 2.4, 2.7). The key structural properties in this pathway are: (a) a flexible, non-bound-like apo receptor interface ensemble that presents an unfavorable binding surface, (b) a core subset of shared ligand binding motifs clustered about an anchor residue that latch the receptor interface but allow it to remain *flexible*, and (e) ligand-specific motifs that split the binding pathway by stabilizing different conformations of the flexible interface.

2.3.6 Molecular triggers could be exploited to design small-molecule PD-1 antagonists.

Bound cocrystal structures of the PD-1 – targeting antibody pembrolizumab reveal that it exploits an evolutionarily-designed induced fit trigger: the four-membered hydrogen bond network that opens Asn66 and makes the receptor interface hydrophobic. This same principle can be applied to design smaller molecular weight PD-1 inhibitors. We have shown that the mGDV motif of the Bristol-Myers-Squibb PD-1 inhibitor combines key pharmacophore features of both PD-L1 and pembrolizumab interfaces: the backbone O of the Gly resembles that of PD-L1's Ala121, the Asp side chain resembles PD-L1's Asp122, and the Val side chain resembles pembrolizumab's Arg102. Simulations suggest that this structural resemblance produces functionally similar dynamics by displacing receptor residue Asn66 (Figure 2.10d) and stabilizing a bound-like, flat hydrophobic surface formed by closed Ile126 and Ile134 (Figure 2.7, 2.11). Docked conformations of the full inhibitor recapitulate most secondary native-like contacts in addition to the core triggering interactions (Figure 2.18, Figure 2.19). Taken together these results support the idea that nature's mechanisms for modulating receptor surfaces might be exploited to design novel chemistries capable of transforming hard to drug targets into more druggable candidates.

2.3.7 Selective promiscuity via induced fit offers potential advantages over conformational selection for multi-ligand regulatory proteins.

Promiscuous regulatory proteins must optimize binding kinetics for multiple ligands by exploiting structural flexibility. Given nature's general mechanisms for flexibility-mediated binding (Figure 2.1), specificity towards multiple ligands could, in principle, be conferred either through conformational selection, by evolving the receptor to intrinsically sample different ligand-specific apo_{BL} states, or by induced fit, by evolving interface interactions that efficiently drive transitions to the ligand-specific EC_{BL} states. If conformational selection is used to achieve multi-ligand specificity, the binding pathway flux will de facto be limited by ΔG_{BL}^{apo} , the free energy difference between each ligand-specific bound-like states

and other states in the apo ensemble. In this scenario, a natural bottleneck would emerge as an increasing number of ligands would lead to lower association rates.

On the other hand, if selective promiscuity is conferred through induced fit, binding pathway flux will not depend on the fractional populations of apo ensemble microstates, but instead will be determined by the ligand-specific triggering mechanisms. We show here that induced fit can efficiently reshape the shallow polar interface of a flexible receptor into a hydrophobic interface amenable to binding multiple ligands by co-evolving a common set of intermolecular contacts. From an evolutionary perspective, this is an efficient approach to spawning novel protein interactions, since these core contacts can be designed just once. Selectivity to novel ligands can then be achieved by evolving relatively small sequence modifications around these core contacts. Perhaps more importantly, we note that contrary to conformational selection, the induced fit approach to selective promiscuity is in principle not limited by the total number of ligands.

It is interesting to note that many well-characterized eukaryotic regulatory domains [46] bind to several linear binding sequences that share common motifs around an anchor residue and differ in other nearby regions. This trend suggests that the selective promiscuity via induced fit mechanism proposed here for PD-1 might apply elsewhere in nature. This possibility is currently being studied by analyzing the triggers of induced fit in other systems.

2.4 METHODS

2.4.1 Initial protein structures used in simulations.

Molecular dynamics simulations (MDs) of the extracellular domain of PD-1 were run in triplicate using the first three solution NMR structures of apo human PD-1 (PDB ID: 2M2D [19]). Before simulating specific receptor - ligand interactions, MDs of apo PD-1 were evaluated to ensure that the resultant dynamics are consistent with the experimentally derived apo NMR ensemble. As shown in Figure 2.12, apo MDs stabilize within about 2.0 Å backbone RMSD of their respective NMR starting points, suggesting that we can successfully sample native-like unbound states.

Available co-crystal structures of human PD-1 / human PD-L1 (PDB ID: 4ZQK [20]), murine PD-1 / human PD-L1 (PDB ID: 3BIK [22]) and murine PD-1 / murine PD-L2 (PDB ID: 3BP5 [21]) complexes were used as templates for placement of peptides in bound-like loci at the receptor interface, and the dynamics of the PD-1 binding interface in response to interactions with different structural motifs on the ligands were analyzed. We focus on interactions relevant for the opening and closing of the pocket around Asn66. Based on co-crystals, we noticed that the core interacting residues of PD-L1 (Ala121, Asp122, Tyr123) and the homologous residues on PD-L2 (Trp110, Asp111, Tyr112) form critical hydrogen bonds (hydrogen bonds) shaping this pocket. Thus, to dissect the contribution of each contact, we simulate the effects of the receptor interacting with a diverse set of peptide derivatives of these specific ligand residues.

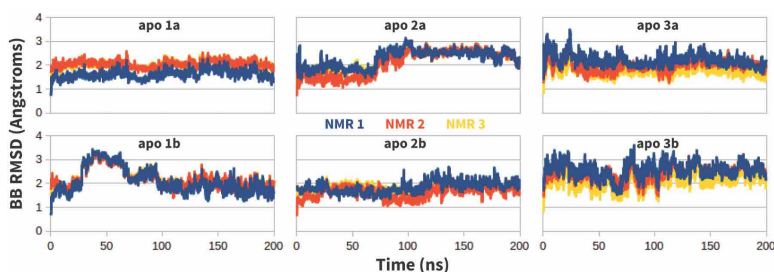


Figure 2.12: Stability of apo PD-1 simulations. Backbone RMSD of apo PD-1 to the first three NMR models (shown in blue, red, and yellow, respectively). Data is shown for six simulations: two replicates (a,b) starting from each of the first three NMR models (1,2,3).

2.4.2 Peptide ligand mimics used in simulations.

Ten distinct PD-1 systems were simulated in order to dissect the ligand groups that trigger induced fit interface deformations on the receptor. These systems included the apo receptor in isolation and in complex with nine different peptides that mimic cognate ligand backbone and side chain interactions with the receptor (Figure 2.3): the anchor residue Tyr, the backbone peptide GGG, five peptides to probe role of ligand side chain contacts DY, GGY, GDG, ADG, GDY, the PD-L1 peptide ADY, and the mGDV peptide, which mimics a patented PD-1 inhibitor.

2.4.3 Simulating PD-1 – peptide interactions.

To generate initial structures for our receptor-peptide MDs, NMR models 1-3 of the human PD-1 were backbone aligned to the murine receptor co-crystal [22] and peptides were modeled after the corresponding human PD-L1 interface residues Ala121 – Tyr123, homologous to PD-L2 interface residues W110 - Tyr112. Systems are simulated for 200 ns, resulting in three replicate MDs per system (including the apo PD-1 system, which does not include any peptide), and receptor interface dynamics are compared across systems to identify the ligand motifs and interactions responsible for structural transitions towards the bound-like receptor state. Harmonic restraints (100.0 kcal/mol) on all heavy atoms of ligand-mimicking peptides were used in simulation to prevent dissociation of the peptide from the receptor interface.

In peptide MDs, harmonic restraints (100.0 kcal/mol) were also placed on backbone atoms of non-interface PD-1 beta sheets residues 50-55, 80-81, 96-98, 106-109 and 120-122. These residues exhibit < 0.35 Å backbone RMSD in the apo NMR ensemble, and previous studies have also shown that the conformational changes induced by ligand binding do not propagate through the major fold of PD-1 [19]. Hence, these restraints should not prevent our ability to sample the native-like binding dynamics of the

receptor interface in biological conditions. The resolved portion of the N-terminal tail of PD-1 (residues 33-36), which in NMR models has $< 0.65 \text{ \AA}$ backbone RMSD, was also restrained so as to limit artificial mobility that might result from the fact that residues 1-32 were missing from simulation.

2.4.4 Encounter complex modeling and simulation.

Human PD-1 – PD-L1 and PD-1 – PD-L2 encounter complexes were modeled and then simulated in triplicate to probe induced fit trajectories and determine the chronology of inter-molecular interactions and specific interface deformations. We modeled encounter complexes by rigid body docking the extracellular domain of the apo receptor and the Ig-like V-type domains of the apo ligands, allowing no structural overlaps. Docked models of PD-L1 had an average backbone RMSD of 5.7 ± 1.2 to the human PD-1 – PD-L1 cocrystal. Docked models of PD-L2 had an average backbone RMSD of 4.8 ± 1.8 to the murine PD-1 – PD-L2 cocrystal (no human cocrystal is currently available for the PD-1 – PD-L2 complex).

Structural models of apo human PD-L1 and PD-L2 that we used when building encounter complexes were generated by simulating the ligands in solution for 400 ns, using a VMD [47] clustering plugin (<https://github.com/luisico/clustering>) to cluster frames by backbone RMSD using a 3 \AA cutoff, and taking the centroid frame of the largest cluster for each ligand. The initial structure for the PD-L1 clustering MDs was taken as the structure of the bound human ligand from the co-crystal complex with murine PD-1 (PDB ID: 3BIK). As there are currently no available crystal structures of human PD-L2, a homology model was built as a starting point for the clustering simulation by manually mutating the bound structure of murine PD-L2 (PDB ID: 3BP5) and minimizing the resulting structure. We used the ClusPro protein-protein docking server [39] to dock the top apo PD-L1 and PD-L2 centroid structures from their respective MDs to the first three NMR structures of apo human PD-1 (all three receptor structures are non-bound-like). Three bound-like candidate models for the PD-1 – PD-L1/2 encounter complexes that correctly anchored Tyr123/112

were chosen from the ClusPro output. We then simulated these encounter complexes for 400 ns to probe the dynamics of the induced fit binding pathway.

2.4.5 Simulation parameters.

We ran MDs using AMBER14's [48] pmemd.cuda module [49] and the AMBER ff12SB force field. The cutoff for non-bonded interactions was set at 10 Å. Systems were simulated in an octahedral TIP3P water box with periodic boundary conditions and a 12 Å buffer around the solute. Cl⁻ ions were added to the solvent to neutralize the charge of the systems. We minimized each system twice and then equilibrated them before beginning production runs. In the first minimization, solute atoms were held fixed through 500 steps of steepest descent and 500 steps of conjugate gradient minimization. In the second minimization only the solute backbone atoms were restrained through 2000 steps of steepest descent and 3000 steps of conjugate gradient. After minimization, system temperatures were raised to 300 K over the course of a 200 ps constant volume simulation (with an integration step of 2 fs) during which the solute was fixed with weak (10.0 kcal/mol) restraints. Bonds involving hydrogens were held at constant length. For the production MDs, the 200 - 400 ns simulations were held at 300 K under constant pressure with the constraints as listed above for each system and an integration step size of 2 fs.

2.4.6 Analysis tools.

The PyMOL Molecular Graphics System v1.7.4.0 was used for structure preparation and analysis [50]. Trajectories were analyzed using VMDv1.9.2 [47] and the MDpocket software package v2.0 [51, 52] for cavity detection and volume / surface area measurement. Measurements of PD-1 binding pocket occlusion, shown in Figures 2.2d and 2.4d,e, were calculated from molecular dynamics simulations of PD-1 using a Python script [46]. Briefly, the script takes a molecular dynamics trajectory and a set of static reference atoms and identifies which reference atoms are overlapped in each frame of the simulation.

Overlap occurs when any simulated atom crosses the “clash radius” of a reference atom, the clash radius being equal to the sum of the van der Waals radii of the two atoms. The output of the script is the fractional occlusion of each reference atom position, equal to the percentage of simulation frames in which that reference atom is overlapped by simulated atoms. This script was used to evaluate the extent to which the Trp110 and Tyr112/123 binding cavities are open in simulations of PD-1 interaction with various peptides, simulations of apo PD-1, and the apo NMR ensemble of PD-1.

2.4.7 Relative free energies of bound-like versus non-bound-like interfaces.

We classified PD-1 interface conformations using two binary order parameters that define whether interface residues Asn66 and Ile126 are in their ‘open’ or ‘closed’ rotamer states. These parameters are used to distinguish the non-bound-like interface, where Asn66 is closed and Ile126 is open, from the PD-L1-specific bound-like state, where Asn66 is open and Ile126 is closed, and the PD-L2-specific bound-like state, where both Asn66 and Ile126 are open (Figure 2.2e). We estimated the energy differences ΔG_{BL}^{apo} and ΔG_{BL}^{EC} (Figure 2.1) using Maxwell-Boltzmann statistics by assessing the bound-like (BL) and non-bound-like (NBL) state population distributions in the apo and encounter complex (EC) receptor ensembles:

$$\frac{\langle n_{BL}^{apo/EC} \rangle}{\langle n_{NBL}^{apo/EC} \rangle} = e^{\frac{-\Delta G_{BL}^{apo/EC}}{k_B T}} \quad (1)$$

In the above equations, $\langle n_{BL}^{apo/EC} \rangle$ and $\langle n_{NBL}^{apo/EC} \rangle$ denote fractional equilibrium populations of the apo / encounter complex receptor ensembles in the bound-like and non-bound-like macrostates, and $k_B T$ is the product of the Boltzmann constant and temperature. We used MDs to generate the equilibrium ensembles of receptor conformations and analyzed the trajectories to calculate $\langle n_{BL/NBL}^{apo/EC} \rangle$ values.

MDs trajectories were analyzed as follows. Reference structures for the open and closed states of Asn66 were defined using its side chain configuration in the first apo NMR model and PD-L1-bound human cocrystal, respectively (Asn66 has $< 0.2 \text{ \AA}$ heavy atom RMSD between PD-L1 and PD-L2 cocrystals 4ZQK and 3BP5). Each frame of the MDs trajectory is labeled with the state to which the simulated Asn66 had the smaller side chain RMSD to the reference structure. Reference structures for the open and closed states of Ile126 were defined using its χ_1 rotamer angle in the murine PD-L2 and human PD-L1 cocrystals, respectively, this angle being the main distinguishing feature between the two different ligand-bound interfaces (Figure 2.2e). Each frame of the MDs was labeled with the state to which the simulated Ile126 had the closest rotamer angle. The free energy changes of opening Asn66 and Ile126 are calculated using eq. (1) and then compared across different simulations in order to identify triggers of interface deformations.

2.5 SUPPLEMENTARY FIGURES

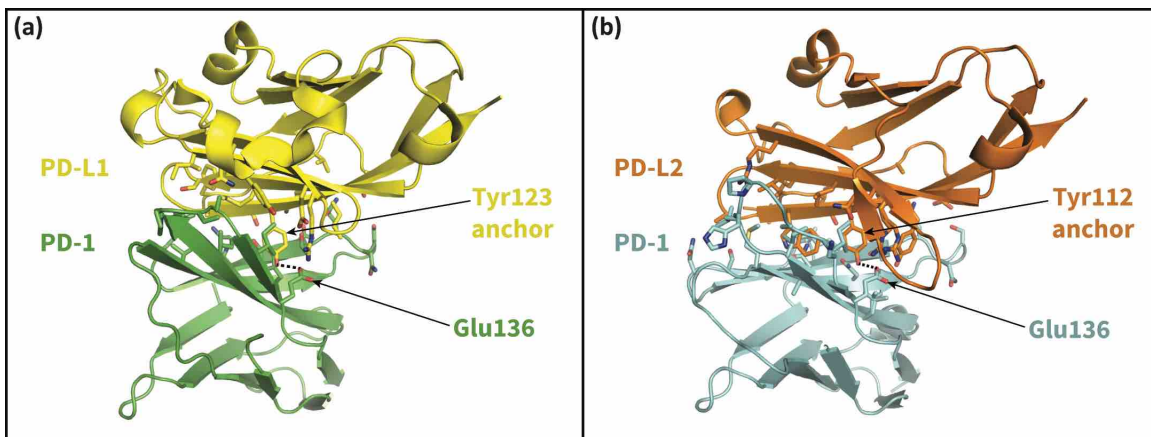


Figure 2.13. The cognate ligands of PD-1. Cocrystal structures of the extracellular domain of PD-1 bound to the Ig-like V-type domains of its two cognate ligands: (a) PD-L1 [20], and (b) PD-L2 [21]. Dashed lines indicate hydrogen bonds between the PD-L1/2 anchor and PD-1 residue Glu136.

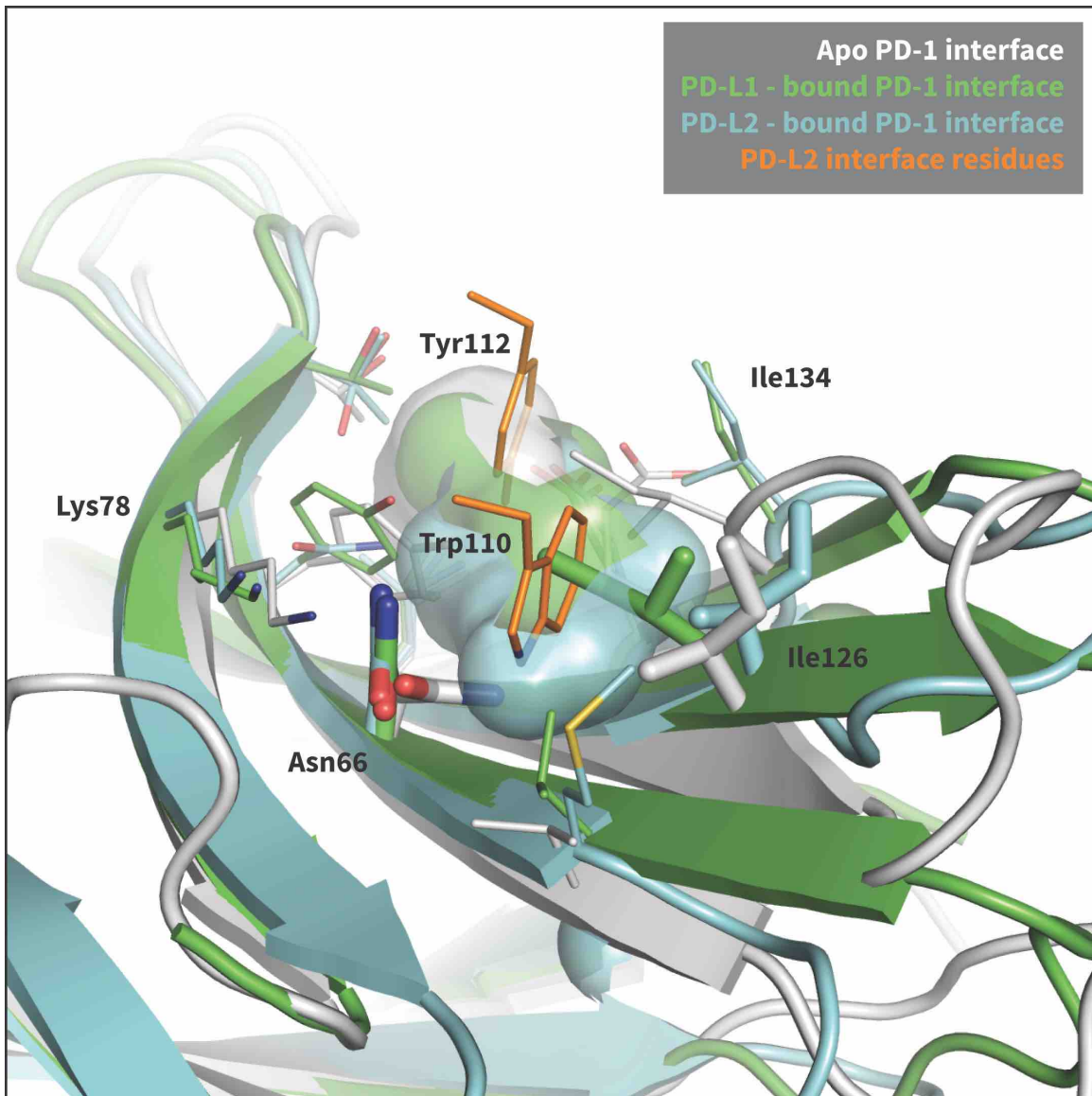


Figure 2.14. Modulation of PD-1's flexible interface cavity. Aligned structures of the apo (white) [19], PD-L1 – bound (green) [20], and PD-L2 – bound (cyan) [21] PD-1 interfaces. Key PD-1 interface residues that line the cavity are shown as small sticks and labelled, with Asn66 and Ile126 shown as large sticks as in Figure 2.2c. The interface cavity volume of each structure is indicated by the transparent surface of matching color. PD-L2 interface residues Trp110 and the conserved Tyr112 anchor are shown as small orange sticks, for reference. The anchor pocket is unstructured in all three receptor states, but only the PD-L2 bound state accommodates Trp110.

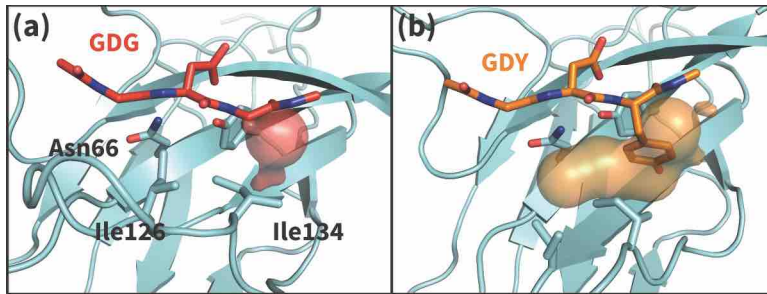


Figure 2.15. Apo PD-1 interactions with GDY peptide opens a hydrophobic cavity. Panels (a) and (b) illustrate the PD-1 interface cavity volume which is plotted in Figure 2.4b. (a) Snapshot from simulation of human PD-1 interacting with the GDG peptide. The PD-1 interface cavity volume is shown in red surface. Although Asn66 is in the open state, the cavity is closed by the closed state of I126. (b) Snapshot from simulation of human PD-1 interacting with the GDY peptide. The PD-1 interface cavity volume is shown in orange surface. The Y anchor side chain positions Ile134 to pull Ile126 out of the pocket via hydrophobic interaction, leaving a large open cavity.

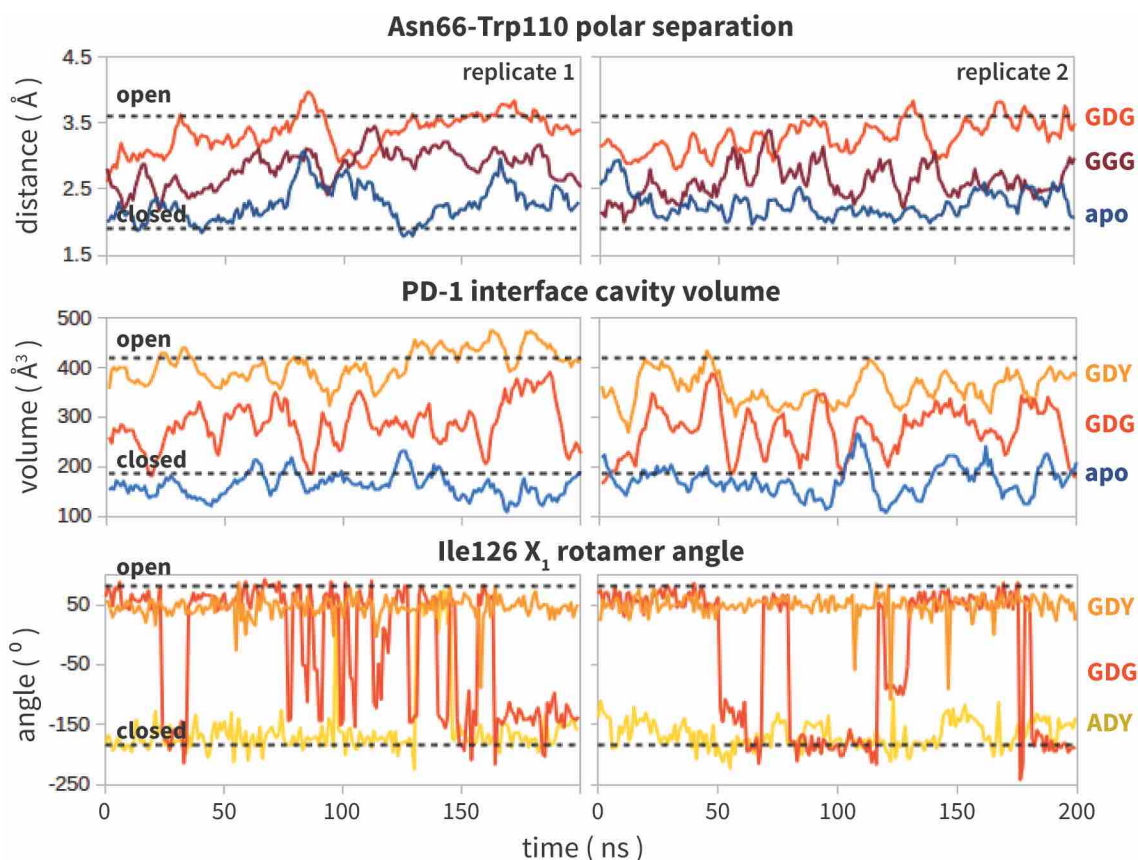


Figure 2.16. Replicate trajectories from Figure 2.4a,b,c. Top: Rolling averages of distance between Trp110_NE1 (from bound PD-L2) and Asn66_ND2 from MDS of apo PD-1 (blue) alone and interacting with GGG (maroon) and GDG (red) peptides. Middle: Rolling averages of PD-1 binding cavity volume from simulations of apo PD-1 alone (blue) and interacting with GDG (red) and GDY (orange) peptides. Bottom: (f) Ile126 X1 rotamer angle from MDS of apo PD-1 interacting with GDG (red), GDY (orange), and ADY (yellow) peptides.

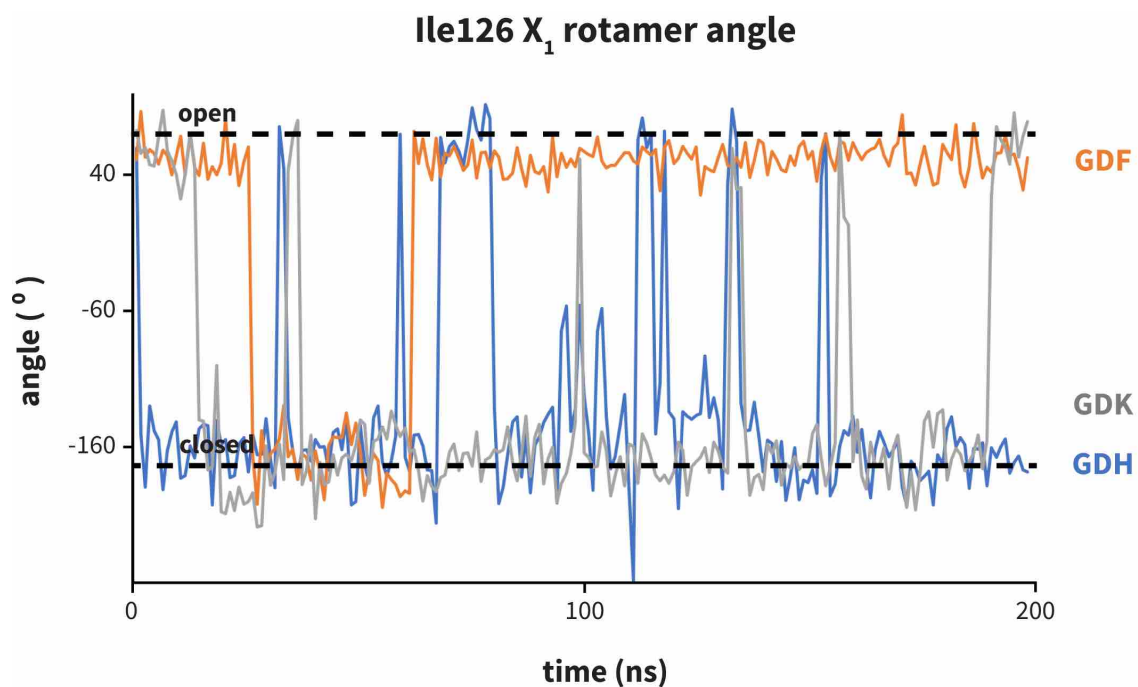


Figure 2.17. Dynamics of PD-1 binding cavity in the presence of different anchor substitutes. Ile126 X₁ rotamer angle from MDs of apo PD-1 interacting with GDF (orange), GDK (grey), and GDH (blue) peptides. GDF produces a mostly open interface cavity, while GDK and GDH stabilize the closed surface.

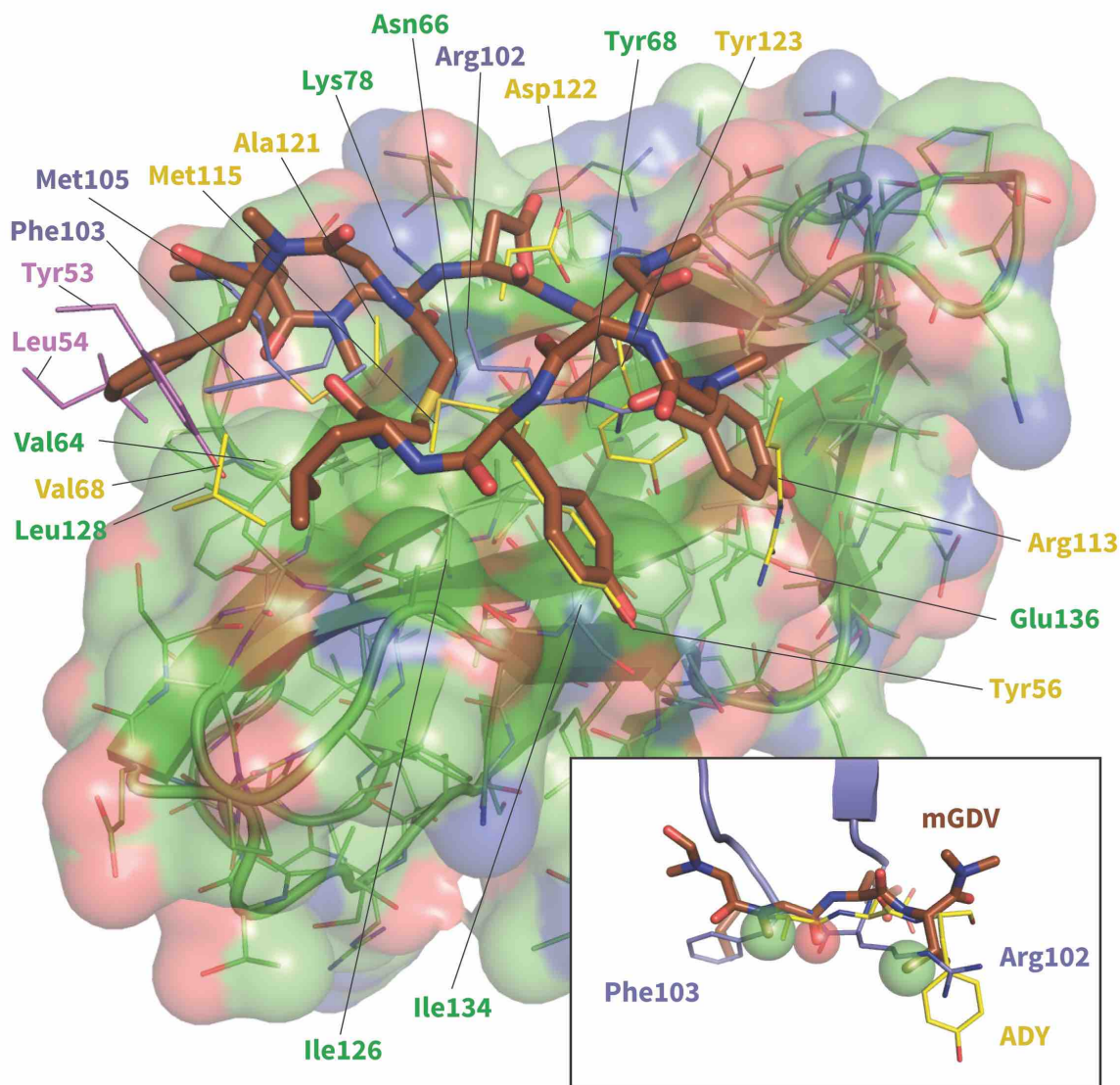


Figure 2.18. Model of potent Bristol-Myers-Squibb macrocyclic PD-1 inhibitor. Predicted macrocycle binding mode is shown brown, with certain side chains omitted for clarity (see Figure 2.19 for full macrocycle structure). Key PD-L1 (yellow) [20] and pembrolizumab (purple and magenta) [40] interface residues from their bound cocrystal structures are shown to highlight predicted native-like contacts. Inset: the mGDV segment of the macrocycle aligned to PD-L1's ADY trigger and pembrolizumab's corresponding interface residues. Green and red spheres represent hydrophobic and polar pharmacophores matched by both pembrolizumab and the mGDV macrocycle motif.

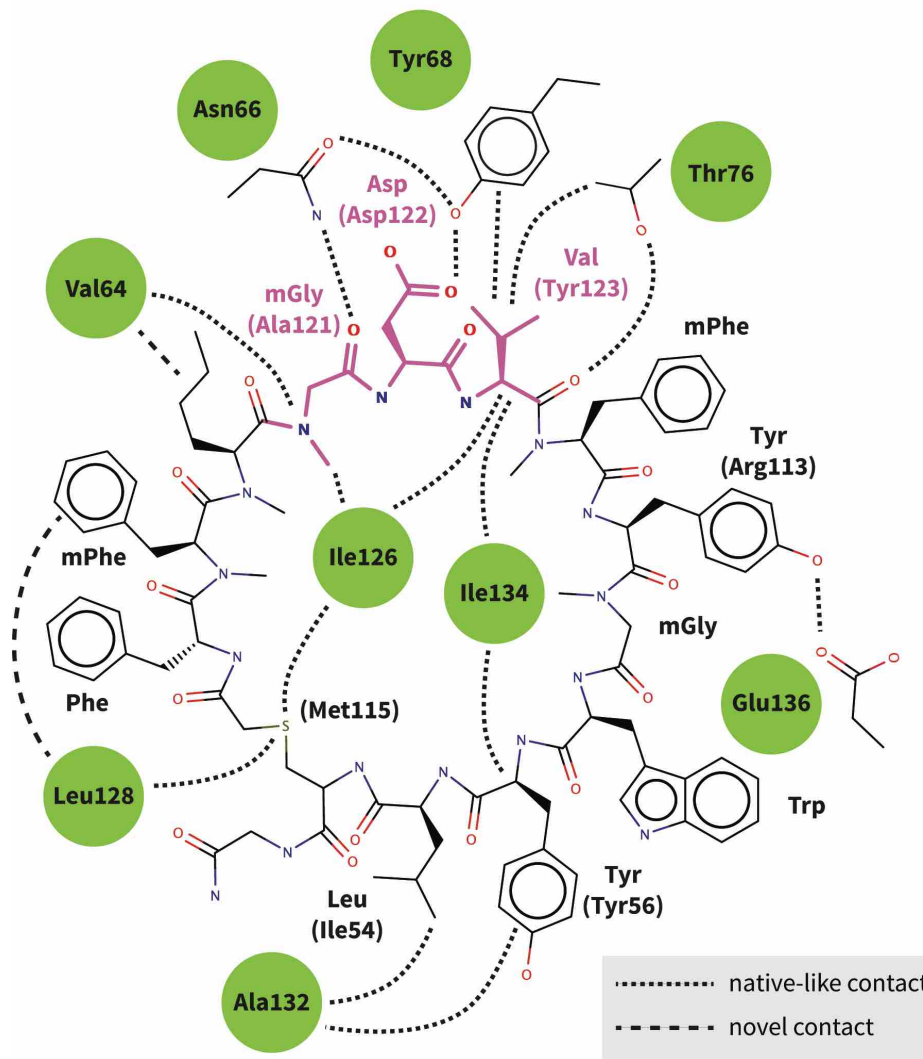


Figure 2.19. Predicted interactions of Bristol-Myers-Squibb macrocyclic PD-1 inhibitor. Figure shows the 2D structure of the patented Bristol-Myers-Squibb macrocycle with the mGDV sequence highlighted in magenta. Dashed lines indicate the specific interactions between the macrocycle and the PD-1 interface (green circles) that are observed in our binding model. Amino-acid components of the macrocycle are labeled, and analogous PD-L1 cocystal [20] residues that participate in the same interactions are indicated in parenthesis. Our binding model recapitulates most native-like contacts present in the human PD-1 – PD-L1 cocystal.

3.0 PREDICTING PROTEIN TARGETS FOR DRUG-LIKE COMPOUNDS USING TRANSCRIPTOMICS

3.1 INTRODUCTION

Most protein research still focuses on roughly 10% of proteins, and this bias has a profound effect on drug discovery, as exemplified by the popular kinase target [53-55]. The origin for this relatively limited exploration of the human interactome and the resulting lack of novel drugs for emerging ‘genomic-era’ targets has been traced back to the availability of small molecular weight probes for only a narrow set of familiar protein families [53]. To break this vicious circle, a new approach is needed that goes beyond known targets and old scaffolds, and that benefits from the vast amount of information we now have on gene expression, protein interactions, their structures and related diseases.

The current target-centric paradigm relies on high-throughput *in-vitro* screening of large compound libraries against a single protein [56]. This approach has been effective for kinases, GPCRs, and proteases, but has produced meager yields for new targets such as protein-protein interactions, which require chemotypes often not found in historical libraries [57, 58]. Moreover, these *in-vitro* biochemical screens often cannot provide any context regarding drug activity in the cell, multi-target effects, or toxicity [59, 60]. On the other hand, the goal of leveraging new chemistries would entail a compound-centric approach that would test compounds directly on thousands of potential targets. In principle, this is regularly done in cell-based phenotypic assays, but it is often unclear how to identify potential molecular targets in these experiments [61-63]. Understanding how cells respond when specific interactions are disrupted is not only

essential for target identification but also for developing therapies that might restore perturbed disease networks to their native states.

Compound-centric computational approaches are now commonly applied to predict drug–target interactions by leveraging existing data. However, many of these methods extrapolate from known chemistry, structural homology, and/or functionally related compounds, and excel in target prediction only when the query compound is chemically or functionally similar to known drugs [64-69]. Other structure-based methods such as molecular docking are able to evaluate novel chemistries, but are limited by the availability of protein structures [70-72], inadequate scoring functions, and excessive computing times, which render structure-based methods ill-suited for genome-wide virtual screening [73].

More recently, a new paradigm for predicting molecular interactions using cellular gene expression profiles has emerged [74-76]. Previous work has shown that distinct inhibitors of the same protein target produce similar transcriptional responses [77]. Related studies have predicted secondary pathways affected by known inhibitors by identifying genes that, when null-mutated, diminish the inhibitory expression signature of drug-treated cells [78]. When no target information is available for the compound in question or related compounds, alternate approaches have mapped drug-induced differential gene expression levels onto known protein interaction network topologies and prioritized potential targets by identifying highly perturbed subnetworks [79-81]. These studies predicted roughly 20% of known targets within the top 100 ranked genes (see Chapter 3.4 for details), but did not predict or validate any previously unknown interactions.

The NIH’s Library of Integrated Cellular Signatures (LINCS) project presents an opportunity to leverage gene expression signatures from other types of cellular perturbations for the purpose of drug-target

interaction prediction. Specifically, the LINCS L1000 dataset contains cellular mRNA signatures from treatments with 20,000+ small molecules and 20,000+ gene over-expression (cDNA) or knockdown (sh-RNA) experiments. Based on the hypothesis that drugs which inhibit their target(s) should yield similar network-level effects to silencing the target gene(s) (Figure 3.1a), we calculated correlations between the expression signatures of thousands of small molecule treatments and gene knockdowns in the same cells. We used the strength of these correlations to rank potential targets for a validation set of 29 FDA-approved drugs tested in the seven most abundant LINCS cell lines. We evaluate both direct signature correlations between drug treatments and knockdowns of their potential targets, as well as indirect signature correlations with knockdowns of proteins up/down-stream of potential targets. We combined these correlation features with additional gene annotation, protein interaction and cell-specific features in a supervised learning framework and use Random Forest (RF) [82, 83] to predict each drug's target, achieving a top 100 target prediction accuracy of 55%, which we show is due primarily to our novel correlation features.

Finally, to filter out false positives and further enrich our predictions, we used molecular docking to evaluate the structural compatibility of the RF-predicted compound–target pairs. This orthogonal analysis significantly improved our prediction accuracy on an expanded validation set of 152 FDA-approved drugs, obtaining top-10 and top-100 accuracies of 26% and 41%, respectively, more than double that of aforementioned previous methods. We applied our pipeline to 1680 small molecules profiled in LINCS and experimentally validated seven potential first-in-class inhibitors for high-impact cancer targets HRAS, KRAS, CHIP, and PDK1.

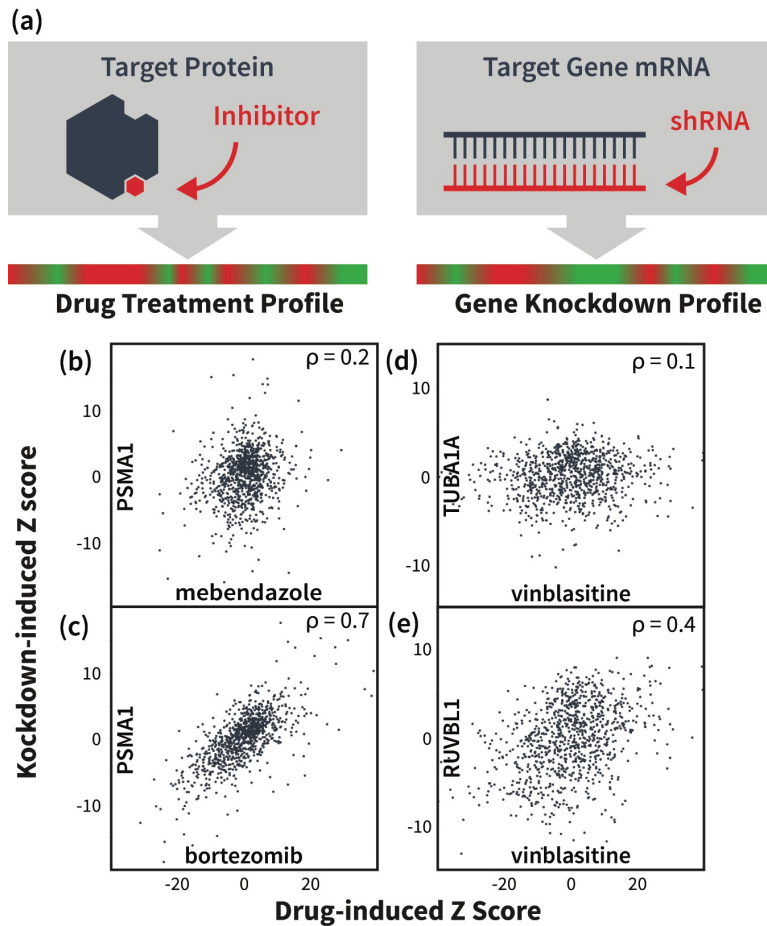


Figure 3.1. Drug and gene knockdown induced mRNA expression profile correlations reveal drug-target interactions. (a) Illustration of our main hypothesis: we expect a drug-induced mRNA signature to correlate with the knockdown signature of the drug’s target gene and/or genes on the same pathway(s). (b,c) mRNA signature from knockdown of proteasome gene PSMA1 does not significantly correlate with signature induced by tubulin-binding drug mebendazole, but shows strong correlation with signature from proteasome inhibitor bortezomib. Data points represent differential expression levels (Z-scores) the 978 landmark genes measured in the LINCS L1000 experiments. (d,e) Signature from tubulin-binding drug vinblastine shows little signature correlation with knockdown of its target TUBA1A, but instead correlates with the knockdown of functionally related genes, such as RUVBL1.

These novel inhibitor candidates validated our hypothesis that drug treatments and target knockdowns cause similar disruptions of cellular protein networks that produce directly correlating differential expression patterns. More interestingly, we discover that these correlations can occur for knockdowns of the drug's actual protein target(s) and/or for genes up/downstream of the target(s). We refer to the latter as "indirect correlations". Several aspects of our approach represent significant step forwards from previous work exploring expression correlations as a means of predicting molecular interactions [84, 85]. Primarily, we do not assume anything about the small molecule or its likely protein target/pathway and our evaluation of both direct and indirect correlations allow us to screen compounds at a much larger scale and with higher accuracy than has been done previously. Furthermore, to our knowledge, this is the first time that pathway connectivity is explicitly considered by indirect correlational effects between drugs and knockdowns of target interaction partners. This approach helps to visualize and quantify the impact of drugs at the cell level and significantly improves in the translational potential of gene expression data to various realms of chemical biology and medicine. Finally, we open source our predictions and methods, providing enriched sets of likely active compounds for hundreds of human targets and presenting a new avenue for identifying suitable (multi-) targets for novel chemistries and accelerating the discovery of chemical probes of protein function.

3.2 RESULTS

3.2.1 Preliminary prediction of drug targets using expression profile correlation features.

We constructed a validation set of 29 FDA-approved drugs that had been tested in at least seven LINCS cells lines, and whose known targets were among 2634 genes knocked down in the same cell lines. For these drugs, we ranked potential targets using the direct correlation between the drug-induced mRNA expression signature and the knockdown-induced signatures of potential targets (Figure 3.1b,c). For each

cell line, the 2634 knockdown signatures were sorted by their Pearson correlation with the expression signature of the drug in that cell line. We used each gene's lowest rank across all cell lines to produce a final ranking of potential targets for the given drug. Using this approach, we predicted known targets in the top 100 potential targets for 8/29 validation compounds (Table 3.1). Indirect correlations were evaluated by the fraction of a potential target's known interaction partners (cf. BioGrid [86]) whose knockdown signatures correlated strongly with the drug-induced signature. Ranking by indirect correlations predicted the known target in the top 100 for 10 of our 29 validation compounds (Table 3.1). Interestingly, several of these compounds showed little correlation with the knockdown of their targets (Figure 3.1d,e), with only 3/10 targets correctly predicted using the direct correlation feature alone.

It is well known that expression profiles vary between cell types [87]. Thus, we constructed a cell selection feature to determine the most "active" cell line, defined as the cell line producing the lowest correlation between the drug-induced signature and the control signature. Ranking by direct correlations within the most active cell line for each drug predicted six known targets in the top 100 (Table 3.1). However, all six of these targets were already predicted by either direct or indirect correlations, strongly suggesting that scanning for the optimal correlation across all cell lines is a better strategy than trying to identify the most relevant cell type by apparent activity.

Table 3.1. Performance of target prediction using different features and methods on the 29 FDA-approved drugs tested in 7 cell lines. DIR: direct correlation feature; IND: indirect correlation feature; CS: cell selection feature; MAX: maximum differential expression feature; MEAN: mean differential expression feature; LR: logistic regression; RF: random forest. Values are for the ranking of the top known target for each drug.

Drug	Random	DIR	IND	CS	MAX	MEAN	LR	RF
vinorelbine	310	126	128	1318	1690	425	28	88
dexamethasone	1498	1891	284	943	315	1143	757	157
dasatinib	2325	1009	94	222	290	2621	182	532
vincristine	1979	473	439	386	2231	2196	456	37
mycophenolate-mofetil	564	1100	1263	2986	100	301	3064	3086
amlodipine	995	1338	2439	1801	1875	974	3037	650
lovastatin	1712	72	811	2078	1124	1068	1334	55
clobetasol	2194	820	21	157	74	15	38	65
calcitriol	2514	1059	2938	221	125	1814	1299	252
flutamide	919	2604	69	2806	463	298	702	647
prednisolone	2382	1439	206	787	402	1068	257	23
nifedipine	940	1225	1465	1285	88	322	3037	2249
vemurafenib	1042	1	82	1	1149	1403	22	2
glibenclamide	29	1415	2028	409	1059	740	1300	366
digoxin	2376	73	1470	118	828	567	732	44
bortezomib	1882	1	1	2	2546	2513	24	5
vinblastine	1612	515	56	100	224	377	38	2
digitoxin	573	89	430	216	521	653	79	50
losartan	645	489	988	770	636	31	735	1931
pitavastatin	1855	1976	1036	1117	90	527	1632	373
digoxin	69	521	776	194	127	559	208	64
hydrocortisone	303	312	72	58	93	122	29	17
paclitaxel	2299	74	121	47	371	1862	79	19
lovastatin	988	1	735	1587	1698	1484	128	100
irinotecan	1742	1023	20	236	128	1886	46	160
vincristine	1394	96	74	17	1272	69	28	9
vinblastine	1359	490	75	1383	373	1735	35	2
raloxifene	2080	2883	1818	1172	1064	479	1114	2520
digoxin	1005	102	1066	112	2096	2027	252	167
Mean Ranking	1365	800.6	724.3	776.9	794.9	1009.6	712.8	471.4
Top 100	2	8	10	6	5	3	11	16

Finally, to incorporate the findings of previous studies that suggest that drug treatments often up/down regulate the expression of their target's interaction partners [79-81], we constructed two features to report directly on the drug-induced differential expression of potential targets' interaction partners. These features compute the maximum and the mean differential expression levels of potential targets' interaction partners in the drug-induced expression profile. The lowest rank of each potential target across all cell lines is used in a final ranking. Though neither expression feature produces top 100 accuracies better than those of our correlation features, maximum differential expression identifies three new targets that were not identified using any of the previous features (Table 3.1).

3.2.2 Combining individual features using random forest (RF).

While each of the features in Table 3.1 performed better than random, combining them further improved results. Using Leave-One-Out Cross Validation (LOOCV) for each drug, logistic regression [83] correctly identified known targets in the top 100 predictions for 11 out of 29 drugs and improved the average known target ranking of all drugs (Table 3.1). However, logistic regression assumes that features are independent, which is not the case for our dataset given the complexity and density of cellular protein interaction networks. Hence, we used RF, which is able to learn more sophisticated decision boundaries [88]. Following the same LOOCV procedure, the RF classifier led to much better results than the baseline logistic regression, correctly finding the target in the top 100 for 16 out of 29 drugs (55%) (Table 3.1). Without further training, we tested the RF approach on the remaining 123 FDA-approved drugs that had been profiled in 4, 5, and 6 different LINCS cell lines, and whose known targets were among 3104 genes knocked down in the same cells. We predicted known targets for 32 drugs (26%) in the top 100 (Additional File 3.1), an encouraging result given the relatively small size of the training set and the expected decline in accuracy as the number of cell lines decreases (Table 3.2).

Table 3.2. Performance of two random forest models on validation set of 152 FDA-approved drugs as a function of cells tested. The number of drugs with targets ranked in top 100/50 are shown for the “on-the-fly” and “two-level” RF classification models. Results are divided into subsets of drugs profiled in different numbers of cell lines. Note that the success rate for RF is significant with $p < 10^{-6}$ based on randomization tests (Figure 3.8).

# of Cells	All	7	6	5	4
# of Drugs	152	29	30	42	51
On-the-fly					
Top 100	58	13	15	16	14
Top 50	42	10	10	12	10
Top 100%	38%	45%	50%	38%	27%
Top 50%	28%	34%	33%	29%	20%
Two-level					
Top 100	63	14	15	22	13
Top 50	54	12	14	20	8
Top 100%	41%	48%	50%	52%	25%
Top 50%	36%	41%	47%	48%	16%

Re-training on the full set of 152 drugs and validating using LOOCV, we tested two alternative RF models: “on-the-fly”, which learns drug-specific classifiers trained on the set of drugs profiled in the same cell types, and “two-level”, which learns a single classifier trained on experiments from all training drugs (see Chapter 3.4 for details). The performances of both methods as a function of the number of cell lines profiled are summarized in Table 3.2. On-the-fly RF correctly ranked the targets of 58 out of 152 drugs in the top 100 (38%), with 42 of them in top 50 (28%). Two-level RF produced better enrichment, correctly predicting targets for 63 drugs in the top 100 (41%), and for 54 drugs in the top 50 (36%). In sharp contrast, random rankings (based on 20000 permutations) leads to only 7% of drugs with targets in the 100, indicating that both our training/testing and LOOCV results are extremely significant (Figure 3.8). It is also

noteworthy that the top-100 accuracy of the two-level RF analysis increases to 50% if we only consider drugs treated in 5 or more cell lines.

3.2.3 Gene ontology analysis of protein targets.

Next, we analyzed in what context our Random Forest analysis was most successful. To do this, we divided the 152 drugs in our training data into “successful” predictions (the 63 drugs for which the correct target was ranked in the top 100), and “unsuccessful” predictions. We also divided the known targets into those that were correctly predicted and those that were not. We considered several different ways to characterize small molecules including molecular weight, solubility, and hydrophobicity, but none of these seemed to significantly correlate with our “successful” and “unsuccessful” classifications. Next, we used gene ontology to test for enrichment of “successful” and “unsuccessful” targets. Interestingly, we found that “successful” targets were significantly associated with intracellular categories, while the “unsuccessful” targets were mostly associated with transmembrane and extracellular categories (Table 3.4).

Based on this result we further incorporated cellular component as a feature in our two-level RF. We encode this feature by assigning 1 to the intracellular genes and -1 to the extracellular ones. We ran the two-level random forest with this additional feature included and demonstrated that the cellular component increases the number of top 100 genes to 66 and top 50 genes to 55.

3.2.4 Structural enrichment of genomic predictions.

Figures 3.1d,e show that the gene regulatory effects of TUBA1A inhibition by the drug vinblastine manifest primarily as indirect correlations with knockdowns of the target’s interaction partners, such as RUVBL1, rather than via direct correlation with knockdown of the target. Such cases reflect the intrinsic

connectivity of cellular signaling networks, which sometimes produce gene expression correlations that are ambiguous with respect to which of the interacting proteins in the affected pathway is the drug's actual target. Our pipeline eliminates some of these false positives using an orthogonal structure-based docking scheme that, although limited to targets with known structure, allows us to significantly improve our prediction accuracy. After performing RF classification on the validation set, we mined the Protein Data Bank (PDB) [89] to generate structural models of the potential targets for our 63 "hits" - drugs for which we correctly identified the known target in the top 100. We selected one or more representative crystal structures for each potential target gene, optimizing for sequence coverage and structural resolution (see Chapter 3.4). We then docked hits to their top 100 potential targets and ranked using a prospectively validated pipeline [90-93].

On average, crystal structures were available for 69 out of the top 100 potential targets for each compound, and structures of known targets were available for 53 of the 63 hits. In order to avoid redocking into cocrystals of our hits, we made sure to exclude from our analysis all crystal structures containing these 53 ligands, ensuring that our results would not depend on prior knowledge of interaction partners or binding modes. As shown in Figure 3.2, molecular docking scores improved the re-ranking of the known target for 40 of the 53 drugs, with a mean and median improvement of 13 and 9, respectively. Based on genomic data alone, the known target was ranked in the top 10 for 40% of the 63 hits. After structural re-ranking, 65% had their known targets in the top 10 candidates, and this value improved to 75% in the subset of 53 drugs with known target structures. These results demonstrate the orthogonality of the genomic and structural screens, showing that molecular docking can efficiently screen false positives in our gene expression-based predictions.

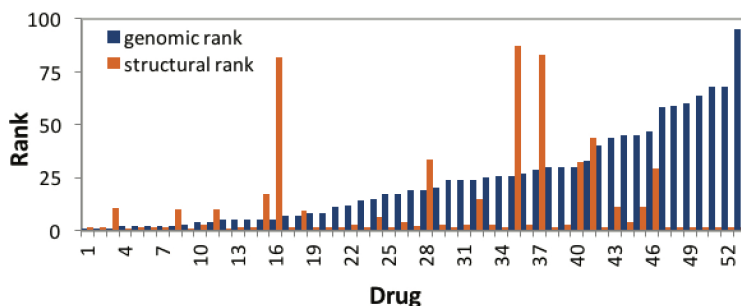


Figure 3.2. Structural enrichment of genomic target predictions. Predicted ranking (lower is better) of the highest-ranking known target for the 53 hits in our validation set with known target structures. Percentile rankings are shown following RF analysis (blue), and following structural re-ranking (orange). Drug names/IDs are listed in Additional File 3.2.

3.2.5 Identifying new interactions in the LINCS dataset.

After validating our approach on known drug targets, we applied our pipeline to a test set of 1680 small molecules and 3333 gene knockdowns and predicted several novel interactions. We applied our pipeline (Figure 3.3) in both compound-centric (target prediction) and target-centric (virtual screening) contexts, in each case producing a final, enriched subset of roughly 10 predictions (either compounds or targets) that we tested experimentally. In compound-centric analyses, we performed molecular docking on the available structures of the input compound's top-100 RF-predicted targets. In target-centric analyses, we ran the RF on our full test set, identified compounds for which the input protein is ranked in the top 100 potential targets, and then docked these candidate inhibitors to the target. In both applications, we analyzed the final docking score distributions and applied a 50% cutoff threshold to identify highly enriched compound/target hits. Structural analysis further facilitated visual validation of the docking models of predicted hits, thereby minimizing false positives.

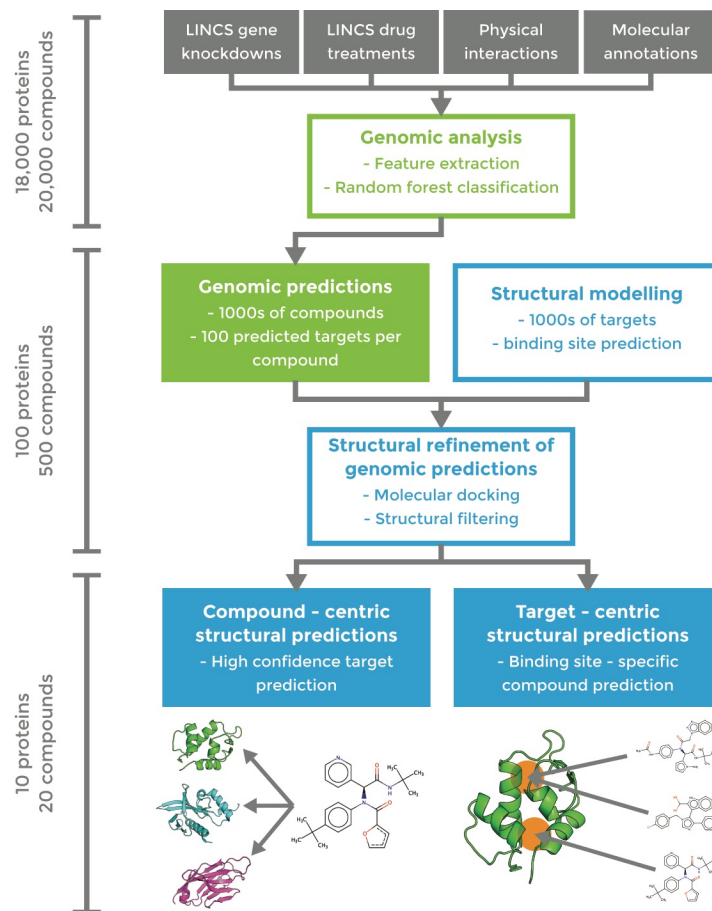


Figure 3.3. Workflow of combined genomic (green) and structural (blue) pipeline for drug-target interaction prediction. Approximate numbers of proteins/compounds in each phase are indicated on the left.

3.2.6 Target-centric prediction of novel RAS inhibitors.

Our first application consisted in identifying novel binders of the high-impact and historically “undruggable” RAS-family oncoproteins. HRAS and KRAS are among the most frequently mutated genes in human cancers [94, 95]. However, despite the extensive structural data available and tremendous efforts to target them with small-molecule therapeutics, as of yet no RAS-targeting drug candidates have shown success in clinical trials [96-98].

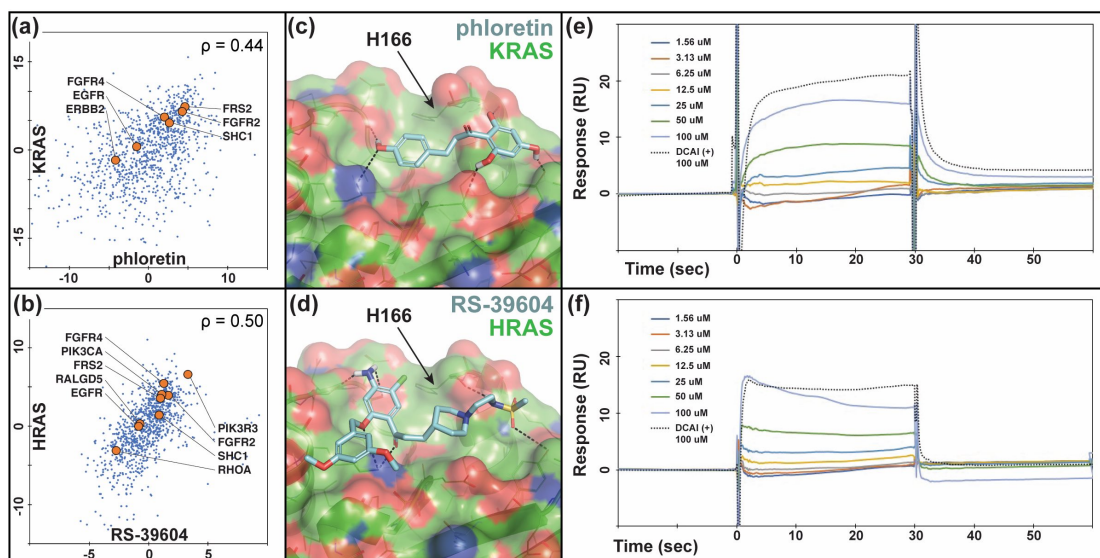


Figure 3.4. HRAS/KRAS inhibitors predicted based on direct correlations and docked poses show direct binding in SPR assays. Differential gene expression profiles of (a) Phloretin and (b) RS-39604 cell treatments and KRAS and HRAS knockdown experiments, respectively. Several functionally related genes listed in BioGrid [86] are indicated to demonstrate the relevance of these profiles as suggestive of direct drug-target interactions. Models of (c) phloretin and (d) RS-39604 bound to an allosteric site on the KRAS and HRAS catalytic domains, respectively. (e) SPR titration response curves for (e) phloretin and (f) RS-39604 binding to KRAS and HRAS, respectively, compared to DCAI positive control.

Among the 1680 compounds in our test set, 84 and 156 were predicted (within the top-100) to target KRAS and HRAS, respectively. These compounds produced mRNA perturbation signatures that correlated strongly with knockdowns of KRAS (Figure 3.4a), and HRAS (Figure 3.4b). Of note, differential expression of genes functionally related to K/HRAS, i.e. FGFR4, FGFR2, FRS1, inform on novel regulatory phenotypes responding to both compound inhibition and gene knock out. We docked predicted compounds to our representative structures of KRAS (PDB ID: 4DSO [96]) and HRAS (PDB ID: 4G0N [99]) (Figure 3.4c,d). RF ranking and docking score distributions were compared to select compounds from our enriched datasets that were both commercially available and moderately priced. Docking models of promising candidates were also examined visually such that to reject models with unmatched hydrogen bonds [100] and select those that showed suitable mechanisms of action (see, e.g., Figure 3.4c,d). We purchased six potential HRAS inhibitors and five potential KRAS inhibitors for experimental validation (Table 3.5).

Our SPR assay measured direct binding of predicted inhibitors to AviTagged HRAS and KRAS. Initial 100 μ M screens showed binding response for compounds RS-3906 against HRAS and phloretin against KRAS, and subsequent titrations confirmed binding at μ M concentrations (Figure 3.4e,f), comparable to the DCAI positive control [96].

3.2.7 Target-centric prediction of novel CHIP inhibitors.

Next, we targeted STUB1, also known as CHIP (the carboxy-terminus of Hsc70 interacting protein), an E3 ubiquitin ligase that manages the turnover of over 60 cellular substrates [101], which to our knowledge lacks specific inhibitors. CHIP interacts with the Hsp70 and Hsp90 molecular chaperones via its TPR motif, which recruits protein substrates and catalyzes their ubiquitination. Thus, treatment with small molecules that inhibit CHIP may prove valuable for pathologies where substrates are prematurely destroyed by the ubiquitin-proteasome system [102].

The screening of the 1680 LINCS small molecules profiled in at least four cell lines predicted 104 compounds with CHIP among the top 100 targets. We docked these molecules to our representative structure of the TPR domain of CHIP (PDB ID: 2C2L [103]), for which we had an available fluorescence polarization (FP) assay. The RF ranking and docking score distributions were compared to select compounds highly enriched in one or both scoring metrics. We next visually examined the docking models of top ranking/scoring hits to select those that show suitable mechanisms of action, and purchased six compounds for testing (Table 3.6). In parallel, we performed a pharmacophore-based virtual screen of the ZINC database [104] using the *ZincPharmer* [93] server, followed by the same structural optimization [90-93] performed on the LINCS compounds. We purchased seven of the resulting ZINC compounds for parallel testing.

Our FP assay measured competition with a natural peptide substrate for the CHIP TPR domain. We found that four (out of six) of our LINCS compounds reliably reduced substrate binding (Figure 3.5a,b), while three (out of seven) ZINC compounds did so to a modest degree (Figure 3.9). The two strongest binders were LINCS compounds 2.1 and 2.2. A functional assay also verified that 2.1 and 2.2 prevented substrate ubiquitination and CHIP autoubiquitination (Figure 3.5c,d, Figure 3.10). Compounds 2.1 and 2.2 also prevented ubiquitination of an alternate substrate that was tested subsequently (Figure 3.11). Importantly, the predicted binding modes of these two compounds did not match the pharmacophore model of the TPR-HSP90 interaction [103], which was used to screen the ZINC database (Figure 3.12). The latter emphasizes the power of our approach to identify novel compounds and mechanisms of action to targets without known inhibitors.

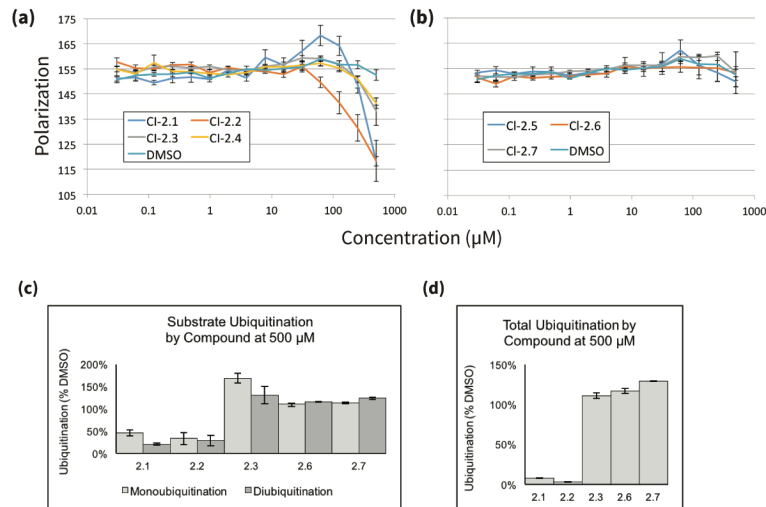


Figure 3.5. Predicted inhibitors show direct binding to and functional inhibition of CHIP. (a,b) Predicted CHIP inhibitors disrupt binding to chaperone peptide by fluorescence polarization. High ranked (a) and low ranked (b) compounds were tested for the ability to compete with a known TPR ligand (5-FAM-GSGPTIEEVD, 0.1 µM) for binding to CHIP (0.5 µM). Results are the average and standard error of the mean of two experiments each performed in triplicate. (c,d) CHIP inhibitors prevent ubiquitination by CHIP in vitro. (c) Quantification of substrate ubiquitination by CHIP from Anti-GST western blot experiments with tested compounds at 500µM, blotted as in Figure 3.10a and normalized to DMSO treated control (2.1, 2.2: N=4; all other compounds: N=2). (d) Quantification of total ubiquitination by CHIP from Anti-GST western blot experiments with tested compounds at 500µM, blotted as in Figure 3.10b and normalized to ubiquitination by a DMSO treated control (all compounds: N=2).

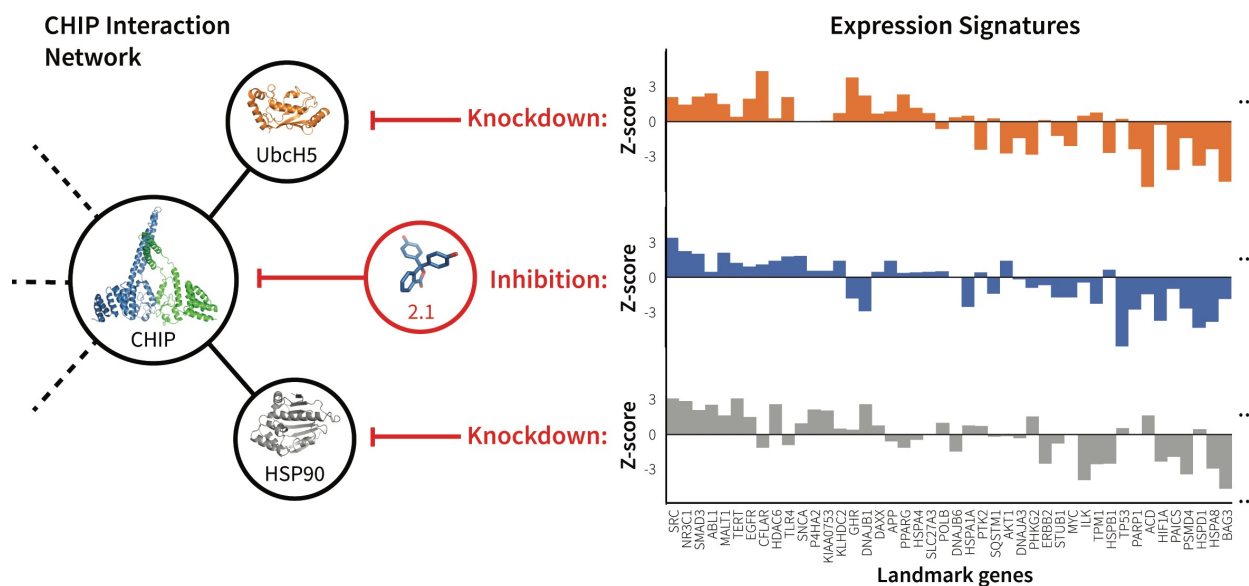


Figure 3.6. mRNA expression signature of CHIP inhibitor 2.1 correlates with knockdown of CHIP interacting partners. The figure illustrates the correlation between the mRNA expression profile signatures produced by treating cells with 2.1 and by knocking down CHIP interaction partners UbcH5 and HSP90. These three perturbations have similar network effects (left), as illustrated by their resulting differential expression signatures (right). For clarity, expression signatures show only the subset of LINCS landmark genes that are functionally related to CHIP according to BioGRID [86].

Contrary to the RAS compounds that were identified based on direct correlations between compound treatments and RAS knockdowns (Figure 3.4a,b), CHIP hits show almost no direct correlation ($\rho_{2.1} = 0.15$, $\rho_{2.2} = 0.02$), but were predicted based on indirect correlations with CHIP interaction partners. Figure 3.6 shows the correlating differential gene expression profiles for compound 2.1 and knockdowns of the CHIP interaction partners UbcH5 and HSP90, which, along with CHIP, were also predicted as potential targets by the RF classifier. However, structural screening ruled out these two partners as potential targets because of a lack of favorable binding modes.

3.2.8 Compound-centric prediction of a novel target for the drug Wortmannin.

We demonstrated a compound-centric application of our pipeline by analyzing Wortmannin, a selective PI3K covalent inhibitor and commonly used cell biological tool. DrugBank [105] lists four known human targets of Wortmannin: PIK3CG, PLK1, PIK3R1, and PIK3CA. Of the 100 targets predicted for Wortmannin, the PDB contained structures for 75, which we used to re-rank these potential targets. Only one known kinase target of Wortmannin, PIK3CA, was detected, and ranked 5th. Our pipeline also ranked 2nd the human kinase PDK1 (PDK1). Although PDK1 is a downstream signaling partner of PI3Ks [106], there is no prior evidence of a direct Wortmannin-PDK1 interaction in the literature. Nevertheless, both the strong direct correlation of wortmannin with the PDK1 knockdown (Figure 3.7a), and the native-like binding mode predicted by our pipeline (Figure 3.7b) suggested a possible interaction.

We experimentally tested this interaction using an alphascreen PDK1 interaction-displacement assay. Since we predicted that Wortmannin binds to the PH domain of PDK1 (Figure 3.7b), we measured the effect of increasing Wortmannin concentrations on the interaction of PDK1 with the second messenger PIP3. We found that Wortmannin specifically increased PDK1-PIP3 interaction, relative to control (Figure 3.7c). Given that PIP3-mediated recruitment of PDK1 to the membrane is thought to play an important regulatory role in the activity of the enzyme [107, 108], a disruptive increase in PDK1-PIP3 interaction following treatment with Wortmannin supports our prediction.

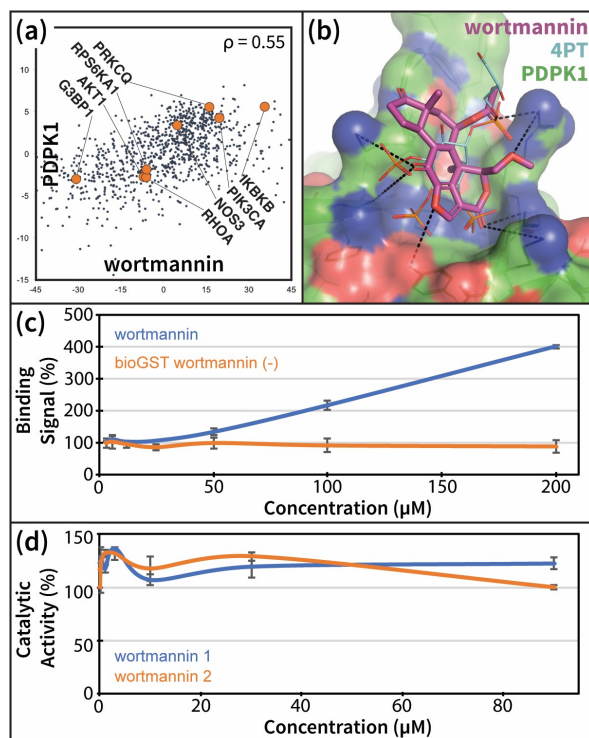


Figure 3.7. Wortmannin promotes PDK1 – PIP3 binding in vitro. (a) Wortmannin treatment and PDK1 knockdown experiments produce directly correlating differential gene expression profiles. Several functionally related genes listed in BioGrid [86] are indicated to demonstrate the relevance of these profiles as suggestive of direct drug-target interactions. (b) Model of wortmannin bound to the PH domain of PDK1, compared to known ligand 4PT (PDB ID: 1W1G [109]). (c) Alphascreen PDK1-PIP3 interaction-displacement assay results for increasing concentrations of wortmannin. Error bars represent the standard error on the mean from two parallel runs. (d) Effect of wortmannin on the in-vitro phosphorylation of the substrate T308tide by the isolated catalytic domain of PDK1. The two lines are from two replicates of the activity assay, with error bars representing the standard error on the mean from two parallel runs for each replicate.

3.2.9 Comparison to existing target prediction methods.

For completeness, we compared results for our 63 hits from the validation set to those produced by available structure and ligand-based methods. HTDocking (HTD) [110] is a structure-based target prediction method that docks and scores the input compound against a manually curated set of 607 human protein structures. For comparison, in our analysis we were able to extract high-quality domain structures for 1245 (40%) of the 3104 potential gene targets. PharmMapper (PHM) [111] is a ligand-based approach that screens the input compound against pharmacophore models generated from publicly available bound drug-target cocrystal structures of 459 human proteins, and then ranks potential targets by the degree to which the input compound matches the binding mode of the cocrystallized ligands. The scope of HTD is limited by the availability of the target structure, while PHM is limited by chemical and structural similarity of active ligands.

HTD and PHM rankings for known targets are shown in Table 3.3, and complete results are shown in Additional File 3.3. Our combined genomics-structure method outperforms the structure-based HTD server (average ranking of the known target is 13 for our method vs. 50 for the HTD server). This suggests that limiting the structural screening to our genomic hits allowed us to predict targets with higher accuracy than docking alone. Results when using the PHM server are on average similar to ours. However, PHM relies on the availability of ligand-bound crystal structures, which in practice makes this class of methods more suitable for drug repurposing than assessing new chemistries or targets.

Table 3.3. Comparison of our pipeline to existing drug-target prediction methods. The average ranking of the highest ranked known target is listed for all 63 validation ‘hits’, for the subset of 53 validation hits with known target structures, and for our seven predicted interactions. ‘Structures available’ indicates the average number of top-100 potential targets with available crystal structures for the compound set. Rankings are compared between the initial random-forest genomic ranking, the structural re-ranking of the top 100 RF predicted targets, the HTDocking server (HTD), and the PharmMapper server (PHM).

	Structures available	Genomic Rank	Structural Re-rank	HTD	PHM
All hits (n=63)	69	22	24	56	23
Hits w/ known target structures (n=53)	71	23	13	50	12
New predictions (n=7)	73	28	31	n/a	n/a

With regards to new validated interactions, alternative approaches failed to predict the interactions with HRAS, KRAS, and CHIP that were verified by our assays. However, a Wortmannin-PDK1 interaction was predicted at the catalytic site by HTD, ranked 540th, and by PHM, ranked 56th. Although we cannot rule out a possible kinase domain interaction, a catalytic activity assay showed that Wortmannin had no measurable effect on the in vitro phosphorylation of the substrate T308tide by the isolated catalytic domain of PDK1 (Figure 3.7d).

3.3 DISCUSSION

Delineating the role of small molecules in perturbing cellular interaction networks in normal and disease states is an important step towards identifying new therapeutic targets and chemistries for drug development. To advance on this goal, we developed a novel target prediction method based on the

hypothesis that drugs that inhibit a given protein should have similar network-level effects to silencing the inhibited gene and/or its up/downstream partners. Using gene expression profiles from knockdown and drug treatment experiments in multiple cell types from the LINCS L1000 dataset, we developed several correlation-based features and combined them in a random forest (RF) model to predict drug-target interactions.

On a validation set of 152 FDA-approved drugs we achieve top-100 target prediction accuracy more than double that of previous approaches that use differential expression alone [80, 81]. Consistent with our underlying hypothesis, the RF results highlight the importance of both direct expression signature correlations between drug treatment and knockdown of the gene target (Figure 3.1c, Figure 3.4a,b, Figure 3.7a) and indirect correlations between the drug and the target's interacting partners (Figure 3.1e, Figure 3.6). Contrary to earlier work [79-81], our method is capable of predicting potential targets for any compound, even those unrelated to known drugs, and our predictions are open source and available for immediate download and testing (<http://sb.cs.cmu.edu/Target2/>). These include potential targets for 1680 LINCS small molecules from among 3000+ different human proteins.

Unlike most available ligand-based prediction methods [64-69], the accuracy of our approach does not rely on chemical similarity between compounds in the training/test sets. For instance, our screen against CHIP, a target with no known small molecule inhibitors, delivered four out of six binding compounds, whereas a parallel analysis using a state-of-the-art structure-based virtual screening [90, 112] yielded only two weak-binding compounds. Moreover, the predicted mechanisms of actions of the more potent LINCS compounds suggest novel interactions that were not prioritized by the ligand-based screen (Figure 3.12).

In contrast to other machine learning methods, our approach reveals important, human-interpretable

insights into perturbation-response properties of cellular networks. Direct and indirect gene expression profile correlations inform on global regulatory responses triggered by small molecule cell treatments (see, e.g., Figures 3.4, 3.6, 3.7). Namely, our genomic screening not only identifies compounds targeting a given protein, but also highlight related genes that are affected by the chemical modulation of the target. This knowledge is bound to play an important role in the design of polypharmacological therapies.

The experimental validation of our predictions for HRAS, KRAS, CHIP and Wortmannin demonstrate the power of our combined genomic and structural pipeline in identifying novel targets and chemotypes. Our prospectively identified modulators are the first of their kind – in that they represent the results of a virtual target-screening process, rather than traditional high-throughput small molecule screening approaches.

Detailed analyses of our predictions suggest several avenues to improve enrichment. We established a clear correlation between the number of cell-types screened and the target prediction accuracy. We identified that a significant source of false positives are indirect correlations that while important to detect the true target, also tend to predict interacting partners as potential targets. Incorporating compound- or target-specific features are also likely to improve our results. For instance, we noticed that our prediction results were less accurate for membrane proteins, and incorporating a cellular localization feature into our RF model increased the number of top-100 hits in our validation set from 63 to 66.

In sum, our method represents a novel application of gene expression data for small molecule–protein interaction prediction, with structural analysis further enriching hits to an unprecedented level in our proteome-scale screens. The success of our proof-of-concept experiments opens the door for a compound-centric drug discovery pipeline that can leverage the relatively small fraction of potentially

bioactive compounds that could be of interest for further investigation to become drugs [113]. Compared to alternative approaches, our method would be particularly suitable for scanning for targets of newly synthesized scaffolds. We are hopeful that our open source method and predictions might be useful to other labs around the world for identifying new drugs for key proteins involved in various diseases and for better understanding the impact of drug modulation of gene expression. Moreover, our approach represents a new framework for extracting robust correlations from intrinsically noisy gene expression data that reflect the underlying connectivity of the cellular interactome.

3.4 METHODS

3.4.1 Data sources

LINCS: LINCS is an NIH program that generates and curates gene expression profiles across multiple cell lines and perturbation types at a massive scale. To date, LINCS has generated millions of gene expression profiles (over 150 gigabytes of data) containing small-molecules and genetic gain- (cDNA) and loss-of-function (sh-RNA) constructs across multiple cell types. Specifically, the LINCS dataset contains experiments profiling the effects of 20,143 small-molecule compounds (including known drugs) and 22,119 genetic constructs for over-expressing or knocking-down genes performed in 18 different cell types selected from diverse lineages which span established cancer cell lines, immortalized (but not transformed) primary cells, and both cycling and quiescent cells.

The gene expression profiles were measured using a bead-based assay termed the L1000 assay¹. To increase throughput and save costs, this assay only profiles a set of 978 so-called “landmark genes” and

¹ <http://support.lincscloud.org/hc/en-us/sections/200437157-L1000-Assay>

the expression values of other genes can be computationally imputed from this set. Note however, that in our analysis we do not rely on such imputation and our methods only need to use the values for the measured genes. In our analysis we used level-4 signature values (containing z-scores for each gene in each experiment based on repeats relative to population control). Data processing of LINCS was done using the l1ktool.²

ChEMBL: To obtain a list of known targets for the drugs in our validation set we used ChEMBL, an open large-scale bioactivity database [114]. We retrieved the records of all FDA-approved drugs using the ChEMBL web service API³. These records contain the designed targets for the drugs along with their synonyms (alternate names) and unique chemical IDs. We used this information to cross-reference these drugs with those in LINCS.

Protein-protein interaction and gene ontology: We obtained PPI information for our feature sets from BioGRID [86] and HPRD [115], both of which contain curated sets of physical and genetic interactions. We retrieved all the records corresponding to protein-protein interactions (PPI) from these data sources and converted them to an adjacency list representation. We obtained the cellular localization of proteins from the Gene Ontology database [116]. We relied on prior analysis [117] to assign the location of for each protein as either “intracellular” (inside of cell) or “extracellular” (outside of cell).

3.4.2 Extracting experiments from LINCS

After determining the subsets of small molecules and cell lines, we obtained the associated experiment identifiers known as “distil IDs” from LINCS meta- information. We included only the reproducible distil

² <http://code.lincscloud.org/>

³ <https://www.ebi.ac.uk/chembl/>

IDs known as “Gold” IDs. We then extracted the corresponding signature values from LINCS using the L1000 Analysis Tools (l1ktools)⁴. We only extracted the signature values of the 978 “landmark” genes because their expression was directly measured, whereas the values of other genes were imputed from the data of these landmark genes.

Drug response experiments

There exist multiple experiments (distil IDs) corresponding to a combination of drug d and cell line c (applying drug d to cell line c). Denote the N_{dc} as the number of experiments for the combination d,c . We extracted a matrix of signature values of size $978 \times N_{dc}$ (number of landmark genes \times number of experiments) per combination. We next took the median of signature values across different experiments, and obtained a 978×1 signature vector per combination. The overall drug-response data Δ , therefore, is implemented as a MATLAB structure with $D = 152$ entries, each containing the following fields.

name: $PertID_d$ (string)
cells: $Cells_{C_d}$ ($|C_d| \times 1$ string array)
signature: $\Delta_{d..}$ ($978 \times |C_d|$)

where $PertID_d$ is the unique internal identifier of a small molecule d in LINCS. $\Delta_{d..}$ contains the expression values of drug d across C_d different cell lines. The $Cells_{C_d}$ field contains cell line names corresponding to the column of $\Delta_{d..}$.

Gene knockdown experiments

We follow a similar protocol to extract the signature values of gene knockdown experiments. Denote N_{gc} as the number of experiments for the combination of gene g and cell line c (knocking down gene g in cell line c). Then, for each combination of g and c we extracted signature values of size $978 \times N_{gc}$. After taking

⁴ <https://github.com/cmapi/l1ktools>

the medians across different experiments, we obtain a 978×1 vector per combination. The overall gene knockdown data Γ has $C = 7$ entries and each entry contains the following fields:

name: $Cells_c$ (string)
genes: $Symbols_{G_c}$ ($|G_c| \times 1$ string array)
signature: $\Gamma_{c..}$ ($978 \times |G_c|$)

where $Cells_c$ is the name of the cell line indexed by c . $\Gamma_{c..}$ contains the signature values of the knockdown of genes in cell line c . The $Symbols_{G_c}$ field is a subset of gene symbols corresponding to the column identifiers of $\Gamma_{c..}$ under the HGNC naming scheme.

Control experiments

We also extracted the signatures of control experiments. The signature values for each cell line were extracted and we obtained a 978×1 vector after taking the medians. We denote the overall control experiment data as Ψ . Ψ is of size $978 \times C$ and implemented with the following format:

name: $Cells_c$ (string)
control: Ψ_c (978×1)

where Ψ_c is the signature column vector for a cell line c .

3.4.3 Building a validation dataset from LINCS

We used ChEMBL to retrieve the reported targets and other meta-information of all FDA-approved drugs, and then cross referenced these drugs with the small molecules profiled in LINCS using their primary product names, synonyms, canonical SMILES strings and standard InChIKey. Based on this analysis we identified 1031 out of approximately 1300 FDA-approved drugs reported in LINCS. However, most of these drugs were profiled in only one or very few cell lines, which meant that relatively little response data was available for them. We thus further reduced this set to 152 drugs profiled in at least 4 cell lines (Table 3.2)

and used these drugs and their known targets as the positive training set. Table 3.9 lists the number of drugs and knockdown experiments available for the seven most abundant cell lines in terms of known targets profiled that we used in our analysis.

3.4.4 Extracting and integrating features from different data sources

The notation and symbols that we use in constructing and using the genomic features are described in Table 3.7 and Table 3.8. Feature construction is summarized below.

Direct correlation: The first feature f_{cor} , computes the correlation between the expression profiles resulting from a gene knockdown and treatment with the small molecule. The correlation feature, denoted as f_{cor} , is constructed as follows:

- For each drug d in Δ ($\Delta_{d..}$):

- Denote T_d as the intersection of gene symbol indices for cells in C_d :

$$T_d = \bigcap_{c \in C_d} G_c$$

- Obtain the knockdown signature values of T_d from Γ . Denote this data matrix as $\Gamma_{C_d \cdot T_d}$, which is of size $|C_d| \times 978 \times |T_d|$, where for each cell line in C_d there is a signature matrix of size $978 \times |T_d|$.

- Compute the Pearson's correlation between $\Delta_{d..}$ ($978 \times |C_d|$) and $\Gamma_{C_d \cdot T_d}$ ($|C_d| \times 978 \times |T_d|$).

Specifically, for each cell line $c \in C_d$, we compute the correlation between $\Delta_{d \cdot c}$ and $\Gamma_{c \cdot T_d}$, and obtain a correlation vector of size $|T_d|$. This is the correlation between the responses of the cells to the drug treatment and their response to the gene knockdown. Each entry in this vector is the correlation of 978 landmark genes of the drug d in one cell line ($\Delta_{d \cdot c}$) and a knockdown of gene g in the same cell

line ($\Gamma_{c.g}$). In other words, if we collect these correlation vectors for all cell lines in C_d and denote the overall correlation feature as f_{cor} :

$$f_{cor}(d, g, c) = corr(\Delta_{d.c}, \Gamma_{c.g}) \quad \forall g \in T_d$$

The correlation feature for one drug d , $f_{cor}(d, \cdot, \cdot)$, has a dimension of $|T_d| \times |C_d|$.

Indirect correlation: Information about protein interaction networks may be informative about additional knockdown experiments that we might expect to be correlated with the small molecule treatment profile. To construct a feature that can utilize this idea we did the following: for each molecule, protein, and cell line we computed $f_{PC}(d, g, c)$, which encodes the fraction of the known binding partners of g (i.e. the proteins interacting with g) in the top X knockdown experiments correlated with this molecule/cell compared to what is expected based on the degree of that protein (the number of interaction partners - this corrects for hub proteins). We used $X = 100$ here, though 50 and 200 gave similar results.

The indirect correlation score is constructed as follows:

- For each drug d in Δ ($\Delta_{d.\cdot}$):
 - Obtain T_d , as defined above.
 - For each cell line c in C_d :
 - Sort T_d in descending order using the correlation values $f_{cor}(d, \cdot, c)$
 - Denote the sorted gene symbol indices for cell line c as $\sigma_c(T_d)$
 - For each knockdown gene g in T_d :
 - Obtain the set of neighbor gene symbol indices from the PPI adjacency list, denote it as N_g .
 - Compute f_{PC} as:

$$f_{PC}(d, g, c) = \frac{|N_g \cap \sigma_c(T_d)_{1:100}|}{|N_g \cap \sigma_c(T_d)| + 50}$$

$f_{PC}(d, g, c)$ has the same dimension as f_{cor} ($|T_d| \times |C_d|$). It reflects the fraction of gene g 's binding partners that are more correlated with drug d in the context of cell line c . We use 50 as the pseudo-count to penalize hub proteins, which have substantially more neighbors than others.

Cell selection: While the correlation feature is computed for all cells, it is likely that most drugs are only active in certain cell types and not others. Since the ability to consider the cellular context is one of the major advantages of our method we added a feature to denote the impact a drug has on a cell line. For each drug/molecule d we compute a cell specific feature, $f_{CS}(d, \cdot)$, which measures the correlation between the response expression profile and the control (WT) experiments for that cell. We expect a smaller correlation if the drug/molecule is active in this cell, and a larger correlation if it is not. The cell selection feature is calculated

Differential expression: In addition to determining the correlation-based rankings of interacting proteins, we also took their drug-induced differential expression into account. We constructed two features that summarize this information for each protein. These features either encode the average or the max (absolute value) expression level of the interaction partners of the potential target protein.

We compute two types of PPI expression scores, denoted as $f_{PE_{max}}$ and $f_{PE_{avg}}$, as follows:

- For each drug d in Δ ($\Delta_{d\cdot}$):
 - For each knockdown gene g in T_d :
 - Obtain N_g , as above (the list of neighbors, or interaction partners, of g)

- For each cell line c in C_d :
 - Find the set of signature values for the neighbors of g , $\Delta_{d,N_g,c}$ (size $|N_g| \times 1$)
 - Compute the two PPI expression scores as:

$$f_{PE_{max}}(d, g, c) = \max(\Delta_{d,N_g,c})$$

$$f_{PE_{avg}}(d, g, c) = \text{avg}(\Delta_{d,N_g,c})$$

Feature data structure

We combined the features for all drugs in a MATLAB structure Ω . Ω has D entries, and each entry $\Omega^{(d)}$ has the following fields:

name: $PertID_d$ (string)
targets: P_d (protein targets for d)
cells: $Cells_{C_d}$ ($|C_d| \times 1$ string array)
genes: T_d (common genes across G_c)
correlation: $f_{cor}(d, \cdot)$ ($|T_d| \times |C_d|$)
PPI correlation: $f_{PC}(d, \cdot)$ ($|T_d| \times |C_d|$)
max PPI expression: $f_{PE_{max}}(d, \cdot)$ ($|T_d| \times |C_d|$)
avg PPI expression: $f_{PE_{avg}}(d, \cdot)$ ($|T_d| \times |C_d|$)
cell selection: $f_{CS}(d, \cdot)$ ($|C_d| \times 1$)

There are a total of $D = 152$ drugs in Ω , and the number of drugs with different values of $|C_d|$ are summarized in Table 3.2.

3.4.5 Subcellular localization assignment

We obtained the cellular localization of genes from the Gene Ontology Consortium. The GO database provides web services to query genes in terms of their associated biological processes, cellular

components and molecular functions in a species-independent manner⁵. We further assign the locations as either “intracellular” (inside of cell) or “extracellular” (outside of cell). The detailed assignments are shown in Table 3.4.

3.4.6 Classification procedure

Criterion of successful classification

Due to the intrinsic noise from the data, we define a successful classification for a drug if any of its correct targets is enriched into the top K ranked genes, where K can be either 50 or 100.

Analysis of feature importance

The evaluation of single features was performed using the drugs that have been applied on all seven cell lines. There are 29 of these drugs from Ω . We sort (descendingly) the common genes T_d for a drug d and cell line c using an individual feature $f(d, ; c)$, where f is either f_{cor} or f_{PC} . Denote $\sigma_d(g, c)$ as the ranking of a gene $g \in T_d$ in the context of cell line c . Then, we define the overall ranking of a gene, $\sigma_d(g)$, to be the best ranking across all seven cell lines: $\sigma_d(g) = \min (\sigma_d(g, c))$ for $c \in C_d$.

Constructing training dataset

Next, we wish to learn and evaluate classifiers that predict drug targets using all features from the feature dataset Ω . We first construct a training data set (design matrix X and its associated labels y) from the feature dataset Ω .

For each drug d in Ω , we select the rows corresponding to the targets in P_d from the other feature matrices and concatenate them into a row vector. The same cell selection vector is appended to every row of

⁵ <http://geneontology.org/page/go-enrichment-analysis>

targets. These rows are assigned with a positive label 1. We then randomly sampled 100 non-target genes (denoted as v_d) and construct the row vectors the same way as the target genes, and these rows are assigned with a negative label 0. In other words, the training matrix and label vector constructed from a drug d are of the following format:

$$X_d = \begin{bmatrix} f_{cor}(d, P_{d1}, \cdot) & f_{PC}(d, P_{d1}, \cdot) & f_{PE_{max}}(d, P_{d1}, \cdot) & f_{PE_{avg}}(d, P_{d1}, \cdot) & f_{CS}(d, \cdot) \\ f_{cor}(d, P_{d2}, \cdot) & f_{PC}(d, P_{d2}, \cdot) & f_{PE_{max}}(d, P_{d2}, \cdot) & f_{PE_{avg}}(d, P_{d2}, \cdot) & f_{CS}(d, \cdot) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{cor}(d, P_{dm}, \cdot) & f_{PC}(d, P_{dm}, \cdot) & f_{PE_{max}}(d, P_{dm}, \cdot) & f_{PE_{avg}}(d, P_{dm}, \cdot) & f_{CS}(d, \cdot) \\ f_{cor}(d, v_{d1}, \cdot) & f_{PC}(d, v_{d1}, \cdot) & f_{PE_{max}}(d, v_{d1}, \cdot) & f_{PE_{avg}}(d, v_{d1}, \cdot) & f_{CS}(d, \cdot) \\ f_{cor}(d, v_{d2}, \cdot) & f_{PC}(d, v_{d2}, \cdot) & f_{PE_{max}}(d, v_{d2}, \cdot) & f_{PE_{avg}}(d, v_{d2}, \cdot) & f_{CS}(d, \cdot) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{cor}(d, v_{d100}, \cdot) & f_{PC}(d, v_{d100}, \cdot) & f_{PE_{max}}(d, v_{d100}, \cdot) & f_{PE_{avg}}(d, v_{d100}, \cdot) & f_{CS}(d, \cdot) \end{bmatrix}; y_d = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where $m = |P_d|$, the total number of targets for drug d . Therefore, the training matrix X_d for drug d is of size $(m + 100) \times 5|C_d|$, and label vector y_d has length $(m + 100)$.

3.4.7 Extending random forests to drugs with missing features

Since our goal here is to predict targets for as many small molecules as possible, we did not want to restrict our analysis to molecules that were only profiled in a large number of cell lines. As noted above, requiring at least seven cell lines reduces the number of known drugs that can be evaluated from 152 to 29 and leads to a similar reduction in the number of novel small molecules that can be evaluated. Thus, it is highly desirable that our classifiers can handle missing data (i.e., cells for which experiments were not performed). To this end, we developed two distinct methods to deal with different compound-specific cell line combinations and extended the random forest [82, 83] model so that can handle molecules profiled in less than seven (but more than four) cell types.

In the first method we simply build the random forest “on-the-fly”. For a given drug i , we iterate through all other drugs in Ω and test if a drug d was profiled in at least all cells which drug i was profiled in. In other words, we test if $C_i \subseteq C_d$ and if so we extract the features of corresponding cell lines in C_i from $\Omega^{(d)}$ and include them in the training data. After we include data for all compatible drugs we can use the training data to train and apply a random forest for the given drug i . We note that for any drug in Ω , there are at least 28 compatible drugs because 29 drugs have been applied to all seven cell lines. However, the main disadvantage of this method is that we need to train separate random forest for every test drug.

In the second method we perform a “two-level” random forest construction process. Here, in addition to the standard step of selecting a (random) subset of the features for each of the trees in the forest we included a step that selected a (random) subset of cells for each of the trees. Specifically, in the first step, we randomly sample four cell lines from the seven total cell lines (denoted as C_i). In the second step, we find all drugs $d \in \Omega$ such that $C_i \subseteq C_d$, extract their features, and use them to train that tree. We repeat this process 3500 times, such that each combination of four cell lines is expected to have roughly 100 trees ($\binom{7}{4} = 35$). To apply this two-level random forest to a test drug t with cell line profile C_t , we select from the forest those decision trees i for which $C_i \subseteq C_t$ and use them to predict the targets for t . Note that unlike the on-the-fly method above, here we only need to train one forest for the entire prediction task.

3.4.8 Generating structural models for docking

In order to use molecular docking to enrich of our random forest predictions, we needed to generate structural models for the genes profiled in LINCS. The union of our top 100 target predictions for the 1680 small molecules profiled in LINCS in at least four cell lines consisted of 3333 unique human genes. We

used a python script (available on github⁶) to mine the PDB for structures of these genes via its RESTful Web Service interface⁷ using Uniprot primary gene name as the search criteria. Crystal structures were available for 1245 of the 3333 human genes in our analysis. The mean and median numbers of structures for these 1245 genes were 11 and 3, respectively. We then analyzed the structures for each gene and selected representative structures that would be used for docking. Representative structure selection was performed automatically using a procedure (explained below) that attempts to optimize sequence coverage, structural resolution, and structural diversity.

To select representative structures, we first divided each gene's structures into "high" and "low" resolution categories using a 2.0 Å threshold. Small structures with less than 20 amino acids were discarded. We then used a greedy algorithm to assess sequence coverage for the remaining structures and select (as representative) the fewest and highest resolution structures that would cover the most of the protein sequence. Redundant structures, defined as structures that did not contain at least 10 residues that were not contained in any of the larger or higher resolution structures, were discarded unless they represented a unique conformation of the protein. Protein conformation was evaluated using ProDy [118] and was considered "unique" if the redundant structure had an all atom RMSD to each of the other representative structures that was above a cutoff threshold that could range between 4.0 Å and 10.0 Å. The specific value of the threshold used for each gene was chosen to try to minimize the number of redundant structures that would be docked against, and higher cutoffs were used for genes that had many redundant structures representing different conformations. After selection, the mean and median numbers of representative structures per gene were 2 and 1, respectively. Each representative structure consisted of exactly one amino acid chain and coordinated ions but without cocrystal ligands or

⁶ https://github.com/npabon/generate_gene_models

⁷ <https://www.rcsb.org/pdb/software/rest.do>

crystallographic waters. We note that this automated procedure is not necessarily tailored to produce representative structures for functional oligomers, since only one chain is considered at a time.

3.4.9 Docking procedure

Compounds were docked to representative structures of their predicted targets with smina [91], using default exhaustiveness and a 6 Å buffer to define the box around each potential binding site. Docked poses across predicted binding sites [119] on a given target were compared and the highest scoring pose of each compound was selected for further analyses [90-93] and comparison to other targets/compounds.

3.4.10 Comparison to previous expression perturbation target prediction methods

Unlike our method which uses both drug-induced and knockdown-induced mRNA expression perturbations, previous target prediction methods analyzed only the drug data within the context of protein interaction networks [80, 81]. As their primary measurement of prediction accuracy, these works generally report the aggregate Area Under the Curve (AUC) of their gene rankings across all validation compounds. The studies mentioned above achieve AUC values of 0.9 and higher in ranking between 11,000 and 18,000 potential gene targets for each compound. To compare these results against our method, we examined the reported AUC curves and calculated the percentage of compounds for which the correct target was ranked within the top 100 potential targets. Both studies achieved top-100 accuracy of 20-21%.

3.4.11 Experimental Assays involving HRAS and KRAS

Surface Plasmon Resonance Spectroscopy: SPR binding experiments were performed on a Biacore S200 instrument (GE, Piscataway, NJ). Neutravidin (Pierce) was coupled to the carboxymethylated dextran surface of a CM5 sensor chip (GE, Piscataway, NJ) using standard amine coupling chemistry to capture

approximately 10,000 RU. Avi-tagged HRAS and KRAS GDP were captured on flows 2, 3 4 with densities of 2450, 2550 and 2960 RU respectively. A titration series of compounds 6, 7, and 12 diluted from 200 – 0.78 μM (seven 2-fold serial dilutions) and compound 15 diluted from 100 – 1.56 μM (six 2-fold serial dilutions) were prepared in 20mM Hepes, 150mM NaCl, 5mM MgCl_2 , 1mM TCEP, 0.01% Tween 20, 5% DMSO, 5 μM GDP, pH 7.4. A positive control for KRAS-GDP of 250 μM DCAI was included. All compounds were injected over all flow cells at 30 $\mu\text{l}/\text{min}$. The data was processed by subtracting binding responses on the reference flow cells, buffer injections and in addition samples were also corrected for DMSO mismatches using a DMSO standard curve.

Protein production: Avi-HRAS(1-189) and Avi-KRAS4b(2-188) were expressed in *E. coli* as His6-MBP-tev-Avi-HRAS(1-189) and His6-MBP-tev- Avi-KRAS4b(2-188), respectively, and purified essentially as previously described [120] for a His6-MBP-tev-fusion protein.

3.4.12 Experimental assays involving CHIP

Materials: Rabbit anti-GST polyclonal antibody conjugated to HRP was purchased from Abcam (ab3416), mouse anti-ubiquitin monoclonal antibody was purchased from Santa Cruz Biotechnology (sc-8017), and horse anti-mouse polyclonal antibody conjugated to HRP was purchased from Cell Signaling Technology (7076S). E2 enzyme UbcH5b and recombinant human ubiquitin were obtained from Boston Biochem (E2-662 and U-100H, respectively).

Protein purifications: His-Ube1, His-CHIP, GST-Hsc70₃₉₅₋₆₄₆, and GST-AT-3 JD were expressed in and purified from *E. coli* BL21(DE3) competent cells (New England Biolabs). Ube1/PET21d was a gift from Dr. Cynthia Wolberger (Addgene plasmid #34965) [121], pET151/D-TOPO CHIP and pGST||2 Hsc70₃₉₅₋₆₄₆ were gifts from Dr. Saurav Misra [122, 123], and pGEX6p1 AT-3 JD was a gift from Dr. Matthew Scaglione [124,

125]. Transformed cultures were incubated in Luria broth with 100 µg/mL ampicillin at 37°C and shaken at 225 rpm until an OD₆₀₀ of 0.3 was attained. Protein expression was then induced with 500µM isopropyl β-D-1-thiogalactopyranoside (IPTG) and cultures were incubated for 24 hrs. at 18°C (15°C for cells expressing GST-Hsc70₃₉₅₋₆₄₆ or GST-AT-3 JD) before the cells were harvested at 5000 rpm for 10 min at 4°C using an F7S-4x1000y rotor for the Sorvall RC-5B Plus Superspeed centrifuge. Cell pellets were stored at -80°C.

Cells harboring His-Ube1 or His-CHIP were thawed and lysed by incubation in lysis buffer (10 mM imidazole, 50 mM NaPO₄ pH 8, 300 mM NaCl, 5 mM 2-mercaptoethanol, 0.25% Triton-100X, 2 mg/mL lysozyme) for 30 min on ice followed by sonication. Purification of Ube1 required addition of protease inhibitors (1% PMSF, 0.2% leupeptin, 0.1% pepstatin A) during lysis and throughout purification. After centrifugation, lysates were applied to Ni-NTA agarose resin (Qiagen), the column was washed with 30 mM imidazole, and proteins were eluted with 200 mM imidazole. Peak fractions containing His-Ube1 were pooled, dialyzed into 20 mM HEPES pH 7.4, 20 mM NaCl, and further purified by anion exchange chromatography over DEAE-Sepharose (GE Healthcare). Bound protein was eluted with a 50-300 mM NaCl gradient. Purified His-Ube1 and His-CHIP were dialyzed into 50 mM HEPES pH 7, 50 mM NaCl, and His-CHIP was further concentrated by centrifugal filtration (Millipore).

Cells harboring GST-Hsc70₃₉₅₋₆₄₆ or GST-AT-3 JD were similarly thawed and lysed by incubation in lysis buffer (50 mM Tris pH 7.5, 150 mM NaCl, 5 mM 2-mercaptoethanol, 0.25% Triton-100X, 2 mg/mL lysozyme, with protease inhibitors) followed by sonication. After centrifugation, lysates were applied to glutathione agarose (Sigma), the column was washed, and proteins were eluted in 6.8 mg/mL reduced glutathione. Peak fractions for each substrate were pooled and dialyzed into 50 mM HEPES pH 7, 50 mM NaCl.

After isolation, the purity of all proteins was verified by SDS-PAGE followed by Coomassie Brilliant Blue staining. Protein concentration was determined by either Bradford (Bio-Rad) or BCA (Thermo Scientific) protein concentration assays. Purified proteins were flash frozen in liquid nitrogen and stored at -80°C.

Fluorescence polarization assay: Fluorescence polarization (FP) studies were carried out as previously described [126]. Briefly, the FP tracer was composed of a peptide derived from Hsp72/HSPA1A (GSGPTIEEVD) that was coupled at the N-terminus to 5-carboxyfluorescein (5-FAM) via an aminohexanoic acid spacer. This tracer ($K_D \sim 0.51 \pm 0.03 \mu\text{M}$) was used in a competition FP format to estimate binding to CHIP. Tracer concentration was $1 \mu\text{M}$, and the CHIP concentration was $0.5 \mu\text{M}$ in a total volume of $20 \mu\text{L}$ in 50 mM HEPES, 10 mM NaCl, 0.01% Triton X-100, pH 7.4. The final DMSO concentration was approximately 1%. After mixing the components, each black 384 well plate (Corning) was covered from light and incubated at room temperature for 30 min. Polarization values were measured at Excitation 485 nm and Emission 530 nm using a Molecular Devices Spectramax M5 plate reader (Sunnyvale, CA). Data were analyzed using GraphPad Prism 6 software.

CHIP in vitro ubiquitination assay: Reactions were initiated by pre-incubating 125 nM Ube1, $1 \mu\text{M}$ Ubch5b, and $200 \mu\text{M}$ ubiquitin for 30 min at 37°C in 50 mM HEPES pH 7.0, 50 mM NaCl, 2 mM ATP, and 4 mM MgCl_2 . In a separate reaction tube, $10 \mu\text{M}$ purified CHIP and up to $500 \mu\text{M}$ compound dissolved in DMSO were combined and incubated for 15 min on ice, followed by the addition of $3 \mu\text{M}$ of either GST-Hsc70₃₉₅₋₆₄₆ or GST-AT-3 JD, which served as substrates for CHIP-dependent ubiquitination. DMSO in these reactions was <5%. After pre-incubation, the ubiquitin-charged E1/E2 mixture was dispensed after which all reactions proceeded for 15 min at 37°C. Reactions were quenched by addition of SDS sample buffer supplemented with 50 mM EDTA, 20 mM DTT. Quenched reactions were resolved by 10% SDS-PAGE,

transferred to nitrocellulose membranes and western blotted with either anti-GST HRP-conjugated antibody to visualize substrate ubiquitination, or anti-ubiquitin primary antibody, followed by an HRP-conjugated secondary antibody to visualize the amount of total ubiquitination. Products were visualized using a Bio-Rad ChemiDoc XRS+ imaging system and quantified using ImageJ software.

3.4.13 Experimental assays involving PDK1

Materials: Soluble biotin-phosphatidylinositol3,4,5-triphosphate, biotin-PIP3, labeled with biotin at sn1-position, was from Echelon Biosciences Inc. Bio-GST, used as a control in the alphascreen system, corresponds to biotinylated GST, (Perkin-Elmer). The peptide substrate T308tide (KTFCGTPEYLAPEVRR; > 75% purity) were synthesized using Pepscan.

PDK1 constructs: PDK1 CD (1-359) and PDK1 PH (360-556) were cloned in pEBG2T vector in frame with GST, expressed in HEK293 by transient transfection and purified using glutathione-sepharose, as described previously for different GST-fusion constructs [127].

Alphascreen interaction assay: The interaction between GST-PDK1 PH (10 nM) and biotin-PIP3 (20 nM) was measured using alphascreen technology (Perkin-Elmer), a bead-based proximity assay. The displacement of the interaction by Wortmannin was performed as previously described for the catalytic domain of PDK1 [128, 129]. Briefly, the assays were performed in a final volume of 25 μ L in white 384-well microtiter plates (Greiner Bio-One), including the interacting partners in a buffer containing 50 mM Tris-HCl pH 7.4, 100 mM NaCl, 2 mM DTT, 0.01% (v/v) Tween-20, 0.1% (w/v) BSA, and the corresponding concentration of the compound (1% final DMSO concentration). 5 μ L of beads (anti-GST conjugated acceptor beads and streptavidin-coated donor beads) at a 20 μ g/ml (microg/ml) were then added to the mixture and after an incubation of 60 minutes, alphascreen counts were measured in an EnVision

Multiplate reader. To set-up the assays, cross-titration experiments were performed, where the concentration of both interacting partners were varied. The concentration of binding partners in the assays were chosen so that both inhibitors and enhancers of the interaction could be identified. Controls using Bio-GST were performed to rule out unspecific effects on the biotin-GST alphascreen interaction assay system.

PDK1 protein kinase activity assay: The in vitro activity of PDK1 was tested using 100-300 ng purified protein, following the transfer of ^{32}P from radiolabelled [$g^{32}\text{P}$]ATP to the polypeptide substrate T308tide at room temperature (22 °C) in a mix containing 50 mM Tris pH 7.5, 0.05 mg/ml BSA, 0.1% β -mercaptoethanol, 10 mM MgCl_2 , 100 μM [$g^{32}\text{P}$]ATP (5-50 cpm/pmol) and 0.003% Brij, as previously performed. [129]

3.5 SUPPLEMENTARY FIGURES

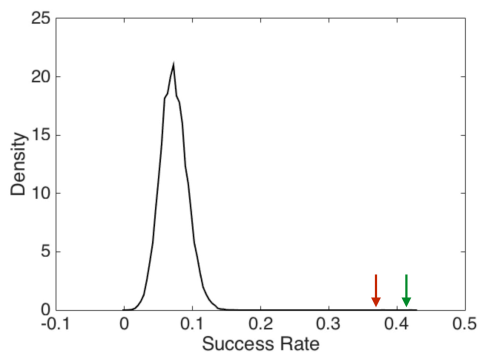


Figure 3.8. Comparing random forest approaches with a random classifier for predicting known targets of validation compounds. The red arrow indicates the success rate of on-the-fly random forest and the green arrow represents the two-level random forest.

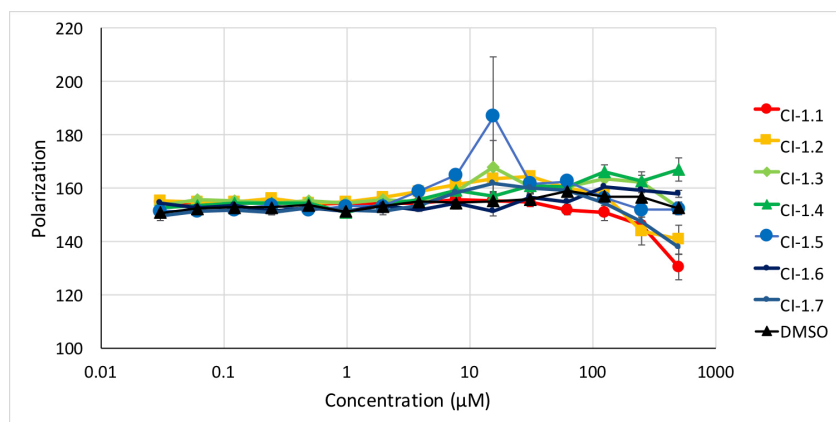


Figure 3.9. ZINC compounds weakly disrupt CHIP binding to chaperone peptide as measured by fluorescence polarization. Results are the average and standard error of the mean of two experiments each performed in triplicate.

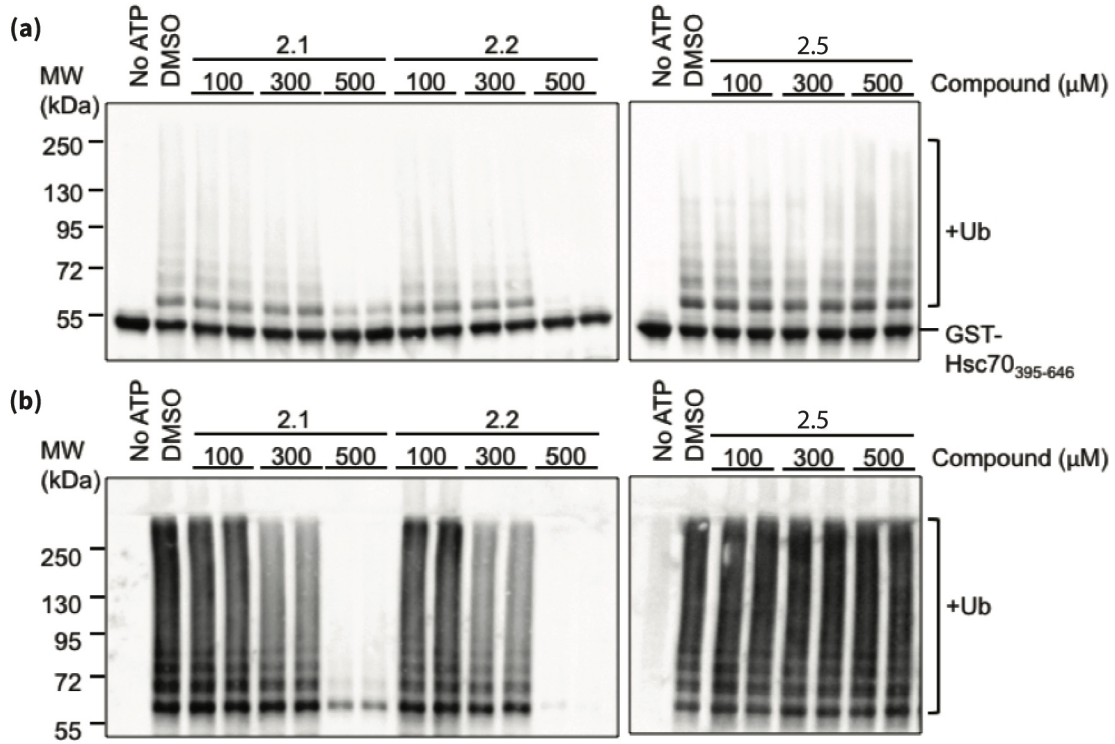


Figure 3.10. CHIP inhibitors prevent ubiquitination by CHIP in vitro. (a) Anti-GST western blot showing substrate ubiquitination by CHIP in reactions treated with high ranked (2.1, 2.2) and low ranked (2.5) compounds. (b) Anti-ubiquitin western blot showing total ubiquitination by CHIP in reactions treated with high ranked (2.1, 2.2) and low ranked (2.5) compounds.

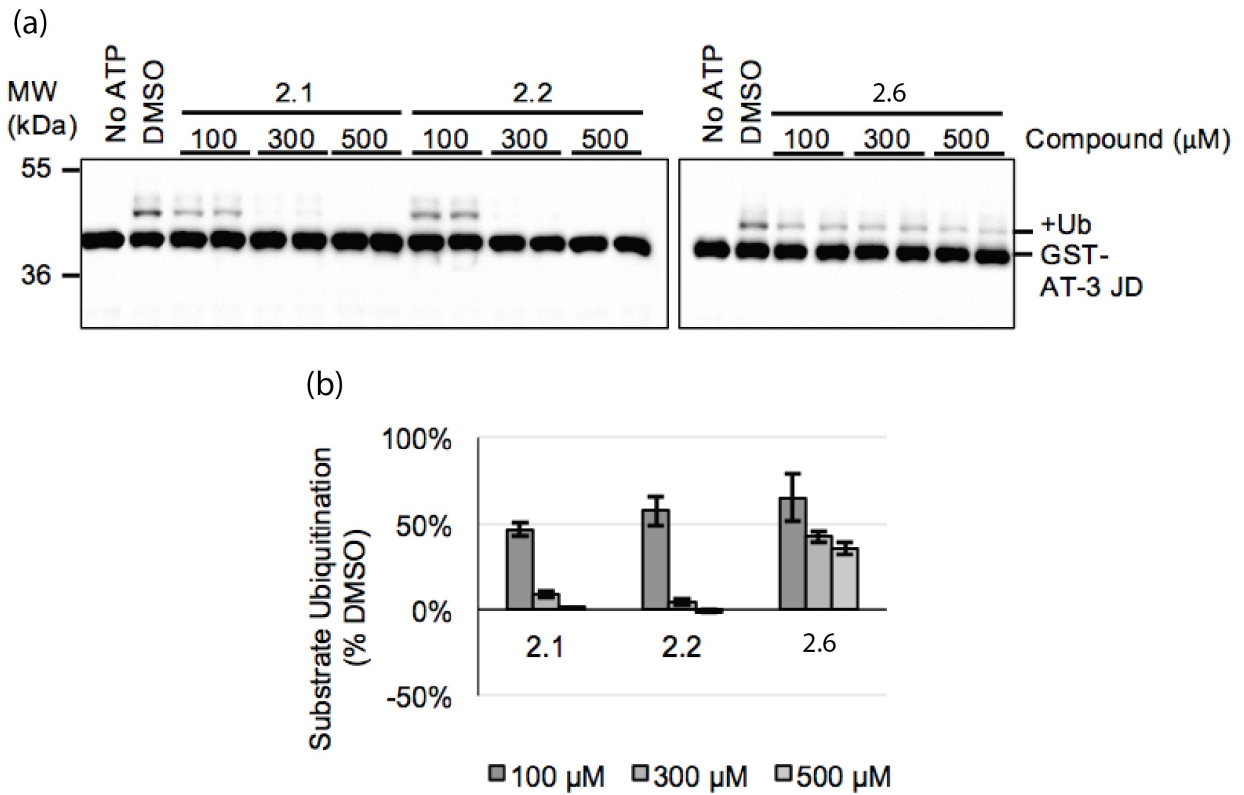


Figure 3.11. Predicted CHIP inhibitors prevent ubiquitination of an alternate substrate. (A) Anti-GST western blot showing AT-3 JD substrate ubiquitination by CHIP in reactions treated with compounds. (B) Quantification of all reactions as in A treated with up to 500 μM compound 2.1, 2.2, or 2.6, normalized to ubiquitination by a DMSO treated control (all compounds: N=4).

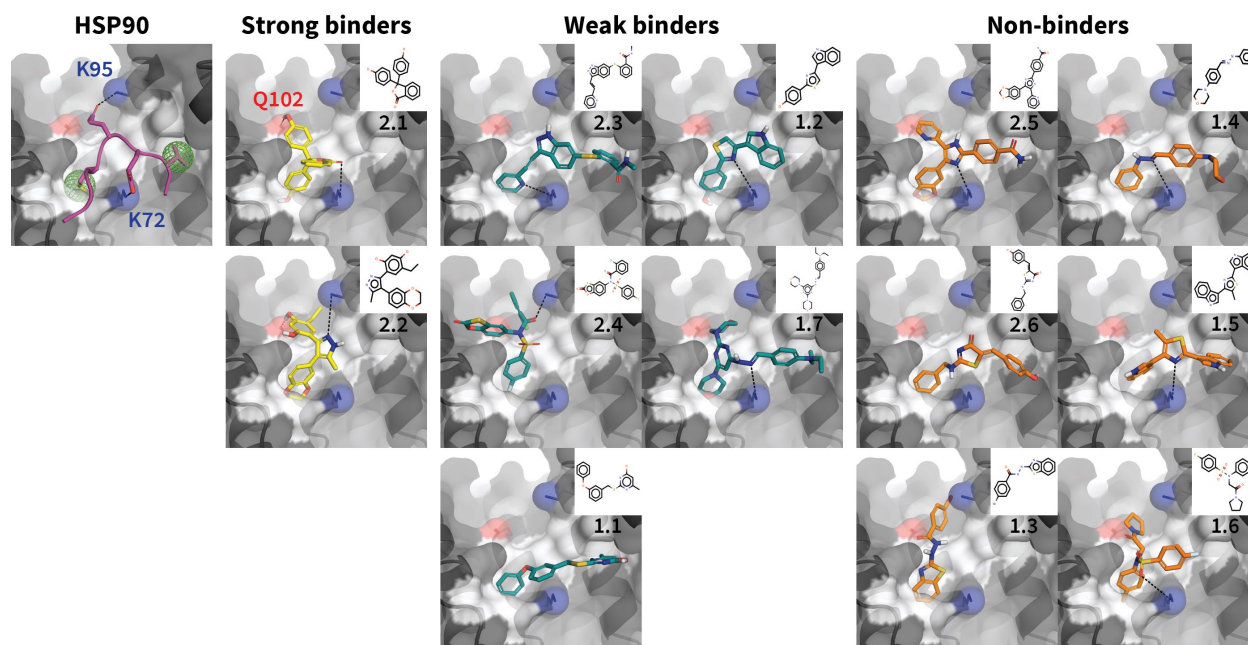


Figure 3.12. Comparison of gene expression-based and pharmacophore-based virtual screens against CHIP. HSP90 shows structure of the CHIP (grey) - HSP90 (magenta) interface (PDB ID: 2C2L [103]), indicating the hydrophobic (green spheres) and polar contact (blue surface / dashed lines) pharmacophores used to screen the ZINC database. **Strong binders** show predicted binding modes for compounds 2.1 and 2.2 from the LINCS screen, which showed the strongest FP signal and robust inhibition of CHIP ligases activity. Interestingly, 2.1 and 2.2 are the only predicted hits to make a novel hydrogen bond to CHIP residue Q102, a contact whose importance is not obvious from the cocrystal structure. **Weak binders** show predicted binding modes for compounds 2.3 and 2.4 from the LINCS screen, and compounds 1.1, 1.2, and 1.7 from the ZINC screen, which showed modest FP signal. **Non-binders** show predicted binding modes for non-binding LINCS compounds 2.5 and 2.6, and non-binding ZINC compounds 1.3 – 1.6.

3.6 SUPPLEMENTARY TABLES

Table 3.4. The cellular localization of successful and unsuccessful drug targets enriched by gene ontology. P-values were computed by intersecting proteins assigned to GO terms listed below with proteins in the sets compared (successful and failed) using the hypergeometric distribution.

	Cellular Component	p-value
Successful Targets	proteasome core complex	7.81E-37
	proteasome core	1.10E-28
	proteasome alpha-subunit	5.68E-18
	cytosol	7.53E-12
	protein complex	1.88E-11
Failed Targets	transmembrane transporter complex	7.77E-15
	sodium-exchanging ATPase complex	4.42E-14
	cation-transporting ATPase complex	8.74E-13
	plasma membrane part	2.19E-11
	chloride channel complex	2.33E-09

Table 3.5. Predicted HRAS/KRAS-targeting compounds purchased for experimental validation. ‘Target Rank’ indicates the ranking of HRAS/KRAS in the RF-predicted list of potential targets for each compound. ‘Cpd Rank’ indicates the structure-based ranking of the compound after docking all candidate inhibitors.

Target	Name	ID	Target Rank	Cpd Rank
HRAS	BRD-A18725729	BRD-A18725729	90	52
	BRD-K00954209	BRD-K00954209	73	1
	BRD-K95858622	BRD-K95858622	92	34
	mefloquine	BRD-K40645748	56	34
	procaterol	BRD-A22684332	99	70
	RS-39604	BRD-K20742498	21	81
KRAS	KM_00799	BRD_K87375115	6	84
	phloretin	BRD_K15563106	65	3
	zardaverine	BRD_K37561857	34	67
	BRD_K85275009	BRD_K85275009	1	80
	amodiaquine	BRD_K91290917	35	32

Table 3.6. Predicted CHIP-targeting compounds purchased for experimental testing. ‘Chip Rank’ indicates the ranking of CHIP in the random-forest predicted list of potential targets for each compound. ‘Cpd Rank’ indicates the structure-based ranking of the compound after docking all candidate inhibitors.

Cpd #	Name	ID	CHIP Rank	Cpd Rank
2.1	phenolphthalein	BRD_K19227686	2	22
2.2	HSP90_inhibitor	BRD_K65503129	2	4
2.3	axitinib	BRD_K29905972	8	13
2.4	BRD_K59556282	BRD_K59556282	11	92
2.5	SB_431542	BRD_K67298865	34	17
2.6	MW_STK33_2B	BRD_K78930611	51	16

Table 3.7. Symbols and notations.

Symbol	Meaning
d	Index for a drug
c	Index for a cell line
g	Index for a gene
N_D	Total number of genes
N_C	Total number of cell lines
C_d	The set of cell line indices for drug d
P_d	The set of protein target indices for drug d
G_c	The set of knockdown gene indices for cell line c
T_d	The intersection of knockdown gene indices G_c for all cell lines in C_d
N_{dc}	Number of experiments for applying drug d to cell line c
N_{gc}	Number of experiments for knocking down gene g in cell line c
N_g	Neighbors, or protein-protein interaction partners, of gene g
Δ	Drug-response data
Γ	Gene-knockdown data
ψ	Control data
Ω	Full feature data
X_d	Training data derived from drug d
y_d	Training label derived from drug d
v_d	Negative (non-target) genes for drug d

Table 3.8. Summary of constructed feature sets. Note that different feature sets can have different dimensions (some contain values for each of the cell lines, etc...). The exact dimension and content of each feature set is discussed in the text.

Feature Name	Symbol	Meaning
Direct Correlation	f_{cor}	Correlation between a drug treatment experiment and a gene knockdown experiment
Indirect Correlation	f_{PC}	Fraction of the known binding partners of a gene in the top X correlated knockdown experiments
Cell Selection	f_{CS}	Correlation between a drug treatment experiment and the control experiment for the cell line
PPI Expression	f_{PE}	The average or the max (absolute value) expression for the known binding partners of a gene

Table 3.9. Cell lines included in the validation dataset. The number of drugs, knockdown genes, and control experiment are shown. For a given cell line, we only include drugs that have their target knockdown experiments available in that cell line.

Cell Line	Drugs	Knockdowns	Controls
A549	188	11947	52
MCF7	180	12031	54
VCAP	175	13225	56
HA1E	172	11968	53
A375	143	11696	58
HCC515	129	7828	52
HT19	96	10185	52

3.7 ADDITIONAL FILES

Additional File 3.1. (additional_file_3.1.xls) Results of testing our random forest classifier on the 123 FDA approved drugs profiled in 4-6 LINCS cell lines after having trained our model on the 29 FDA approved drugs profiled in all 7 LINCS cell lines. The rank of the highest-ranking known target for each compound is listed next to their LINCS ID. We achieve top-100 predictions for 32 drugs, a 26% success rate.

Additional File 3.2. (additional_file_3.2.xls) The names and LINCS IDs of the validation compounds shown in Figure 3.2.

Additional File 3.3. (additional_file_3.3.xls) Structural enrichment of random forest predictions for validation hits and comparison with existing methods. Table lists the 63 'hits' from our validation drug set, including their names, LINCS ID and the number of top-100 predicted targets that had structures available in the PDB. The ranking of the known targets for each compound are shown after our genomic random forest target prediction (GEN), and after our structural re-ranking (STR), along with the percentile rankings produced by alternative target prediction methods HTDocking (HTD) and PharmMapper (PHM). STR, HTD, and PHM values of 100 indicate that the structure of the known target either is not known or was not included in the set of potential targets used by the method.

4.0 DRUGGING THE TNF-INDUCED NF- κ B SIGNALING NETWORK

4.1 INTRODUCTION

The nuclear Factor κ B (NF- κ B) transcription factor regulates expression for hundreds of genes that mediate signals for inflammation, proliferation, and survival [130-135]. Deregulation of NF- κ B has been linked to chronic inflammation in addition to development and progression of various cancers [136-139]. As a pleiotropic regulator of disease-related genes, chemicals that modulate the NF- κ B signaling pathway have therapeutic relevance. However, the complexity of this pathway makes it difficult to understand the mode of action and side effects of these agents. Traditional 'target-centric' drug development strategies that prioritize single-target potency *in-vitro*, and the difficulty of modulating specific protein-protein interactions *in-vivo*, exacerbates the challenges of drugging this pathway in the cell [140]. Not surprisingly, there are no clinically approved inhibitors of NF- κ B pathway components.

An alternative approach is a network-centric strategy to identify small-molecules that inhibit a signaling pathway. Tumor necrosis factor (TNF) is an inflammatory cytokine that initiates dynamic intracellular signals when bound to its cognate TNF receptor (TNFR1). In response to TNF, the I κ B-kinase (IKK) complex is rapidly recruited from the cytoplasm to poly-ubiquitin scaffolds near the ligated receptor where it is activated through induced proximity with its regulatory kinase, TAK1 (Figure 4.1a) [141-146]. When phosphorylated by activated IKKs, NF- κ B inhibitor proteins (I κ B) are degraded and NF- κ B accumulates in the nucleus to regulate transcription. Because the components of the molecular network are well-defined,

disruptors of the signaling network can be predicted from transcriptomic alterations that are shared by i) exposure to small molecules, and ii) genetic knockdowns of pathway components. Through structural screening and live-cell experiments that monitor signaling dynamics in single cells, the dominant mode of action on the signaling network can be inferred.

To meet this challenge and demonstrate a network-centric strategy for targeting TNF-induced NF- κ B signaling, we focused on differential gene expression signatures from the NIH Library of Integrated Network-Based Cellular Signatures (LINCS) L1000 dataset [147]. We compared transcriptional profiles between genetic knockdowns of proteins in the NF- κ B signaling pathway and responses of the same cell types to thousands of distinct bioactive compounds. Using a random forest classification model, we identified compounds whose transcriptomic perturbations resembled genetic disruption. For each compound, the probability of a compound-protein interaction was evaluated in terms of 'direct' correlation with the knockdown signatures, and 'indirect' correlations with knockdown signatures of other proteins in the network. Correlations that cluster on specific protein subnetworks (Figure 4.1a) suggest chemical inhibition within the signaling pathway that mimic genetic inhibition of those network components (Figure 4.4). Note that because of the connectivity of the protein interaction network, it is difficult to precisely identify a single binding target from the gene profiles alone. For example, a compound that disrupts TRADD or TRAF2 might have similar signatures relative to the knockdown of upstream and downstream genes in the pathway such as TNFR1, UBC, or NEMO. Hence, the genomic screening would be expected to show all of these as targets with some probability.

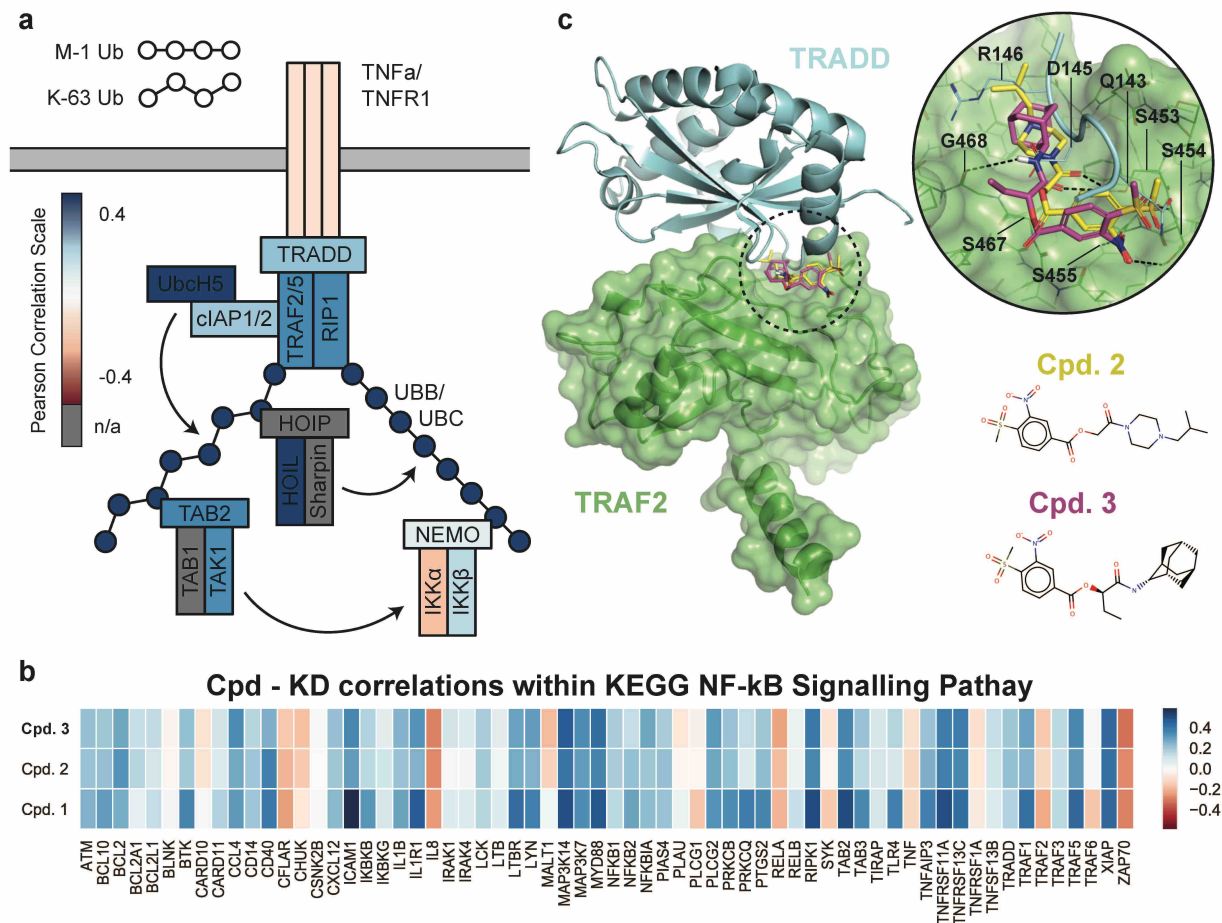


Figure 4.1. Small molecule treatments produce transcriptional responses in that correlate with genetic knockdowns of proteins involved in NF-κB signaling. (a) Schematic of the cytoplasmic multi-protein complex that assembles following ligation of TNF to TNFR1. The color for each protein species in the complex is the average Pearson correlation between gene expression profiles for the species' genetic knockdown and the transcriptional response to compounds 2 and 3. **(b)** Correlation between transcriptomic perturbations by compounds 1, 2, and 3 and the knockdown of genes functionally involved in NF-κB according to the KEGG PATHWAY Database. **(c)** Unbiased molecular docking predicts binding of compounds 2 (yellow) and 3 (magenta) to the TRADD-binding interface of TRAF2. Hydrogen bonds with key TRAF2 interface residues are indicated by dotted lines.

4.2 RESULTS

To identify compounds that inhibit TNF-induced NF- κ B signaling we focused on disruptions of the TNFR1 complex at the level of TRADD, TRAF2, and RIP1, interacting proteins that are necessary to form ubiquitin scaffolding upon TNF stimulation (Figure 4.1a). We reasoned that these disruptions will restrict the dynamics of IKK recruitment and effectively prevent transcriptional signals encoded through nuclear translocation of NF- κ B [148]. Transcriptional signatures for more than 860 compounds showed strong correlations with knockdown of TRADD, TRAF2, and RIP1, so we focused on three compounds (1: BRD-K43131268; 2: BRD-K95352812; and 3: BRD-A09719808) that also correlate broadly with NF- κ B signaling in the KEGG pathway database (Figure 4.1b). For compounds 1, 2, and 3 respectively, predicted targets from our dataset include: TRAF2, UBC, NFKB1, and RIP1; TRAF6, NEMO, TRAF2, NFKB1, UBC, TAB2, and IKK β ; and, NFKB1, TRAF2, UBC, UBB and NEMO. Furthermore, compounds 2 and 3 both showed significant correlations with HOIL, TAK1, cIAP1/2 and UbcH5 knockdowns (Figure 4.1a). Compounds 2 and 3 also had similar chemical structures (Figure 4.1c) and transcriptional profiles (Figure 4.1b), strongly suggesting a similar mechanism of action.

To better predict the likely target of these compounds, we performed molecular docking on available structures/domains in the pathway. The only target suggested by our screening of available structures was TRAF2. The predicted binding modes of all three compounds corresponds to the same binding site as that of TNFR2 (PDB code 1CA9 [149]) and TRADD (PDB code 1F3V [150]) with TRAF2. Compounds 2 and 3 formed hydrogen bond contacts with TRAF2 residues S453, S454, S455, and S467, which are predicted to compete with TRADD interface residues Q143, D145, and R146 based on the co-crystal (Figure 4.1c). Compound 3 is predicted to be a stronger binder due to the extra hydrogen bond formed by its amide group with TRAF2 residue G468. Competitive binding should disrupt the native TRADD-TRAF2 interface and could prevent maturation of the full TNFR1 signaling complex by promoting dissociation or allosteric

stabilization of a non-native conformation. The predicted binding mode of compound 1 is less specific and did not form any of the aforementioned contacts (Figure 4.5). Thermal shift assays showed that compounds 2 and 3 respectively exert a subtle to moderate dose-dependent stabilizing effect on full length TRAF2 (Figure 4.2a, b), suggestive of direct binding, whereas compound 1 did not show a clear trend (Figure 4.6). We note that the observed thermal shifts are consistent with the relatively small stability effect that the compounds are expected to exert on the stable trimer formed by the soluble full length TRAF2 protein [149].

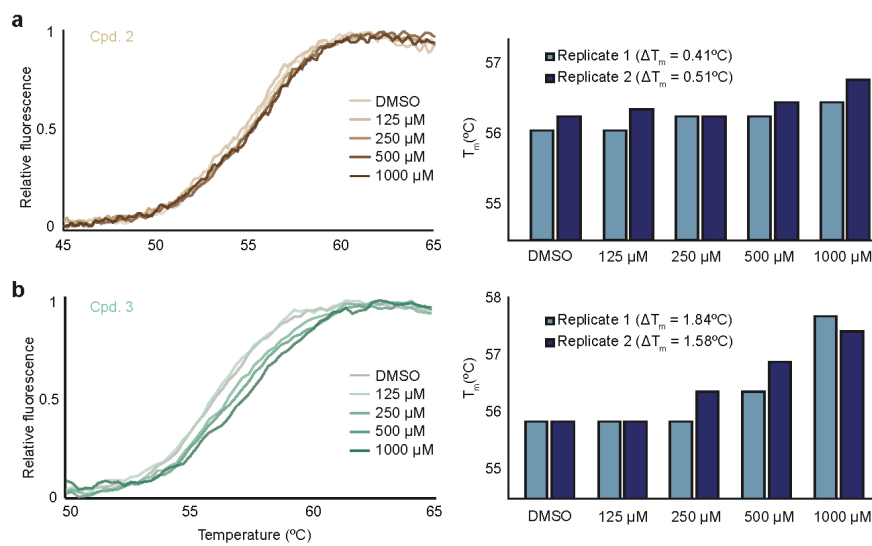


Figure 4.2. Thermal shift assays indicate moderate dose-dependent stabilization of TRAF2 by compounds 2 and 3. Normalized melt curves (left) and melting temperature (Δt_m ; right) of full length TRAF2 were recorded in the presence of DMSO or indicated concentrations of (a) compound 2 and (b) compound 3. The rightward shift of the melt curve in the presence of compounds, quantified by the Δt_m in replicate experiments, suggest increased thermal stability of the protein-compound complex.

We set out to determine whether the compounds are effective inhibitors of NF- κ B signaling in living cells. For this, the endogenous gene locus for the transcriptionally active RelA subunit of NF- κ B was modified using CRISPR/Cas9 to encode a fluorescent protein (FP) fusion in U2OS cells (Figure 4.7), a cell line that forms IKK-recruiting polyubiquitin scaffolds in response to TNF [151]. Responses of single cells exposed to TNF showed transient and variable translocation of NF- κ B into the nucleus when measured from time-lapse images (Figure 4.3a), comparable with other human cancer cell lines that express FP-RelA fusions [148, 152, 153]. When cells were pre-treated with compounds 2 and 3 before exposure to TNF, nuclear mobilization of NF- κ B was reduced in proportion with the concentration of the inhibitory compound (Figure 4.3b). To quantify the compounds' effect on NF- κ B dynamics, each single-cell trajectory was decomposed into a series of descriptors (Figure 4.3c) that transmit information within the cell about extracellular TNF [152]. Although some descriptors did not show a clear trend (Figure 4.8), the most informative descriptor (area under the fold change curve, or 'AUC') was significantly reduced by 1 μ M pretreatment with either compound before exposure to TNF, and most descriptors were nearly indiscernible from untreated control cells when pretreated with 10 μ M (Figure 4.3d). By contrast, compound 1 did not significantly alter the TNF-induced dynamics of nuclear NF- κ B (Figure 4.9). These data suggest that compounds 2 and 3 target the upstream TNFR1 multi-protein complex in the signaling network to restrict NF- κ B activation.

To test this hypothesis, and directly observe the recruitment of IKK to the TNFR1 complex, we used CRISPR/Cas9 to target the γ -subunit of the IKK complex (also known as NEMO) for FP fusion in U2OS cells (Figure 4.10). FP-IKK was diffuse within the cytoplasm and rapidly localized to punctate structures near the plasma membrane after exposure to TNF (Figure 4.3e). The number of punctate FP-IKK structures in single cells peaked at 15 minutes and dissolved within an hour of TNF stimulation (Figure 4.3f). Although the recruitment and dissolution dynamics of FP-IKK are prolonged when compared with a previous study

that overexpressed a fusion of mouse IKK γ [151], they are otherwise qualitatively similar. Consistent with our observations for NF- κ B, the number of TNF-induced puncta were greatly reduced in single cells that were pretreated with compounds 2 or 3 before exposure to TNF (Figure 4.3f). Unexpectedly, the compounds also reduced the overall expression level of IKK γ (Figure 4.11) through an unknown mechanism that may relate to TRAF-dependent ubiquitination cascades that regulate the ambient stability of other NF- κ B-inducing kinases [154].

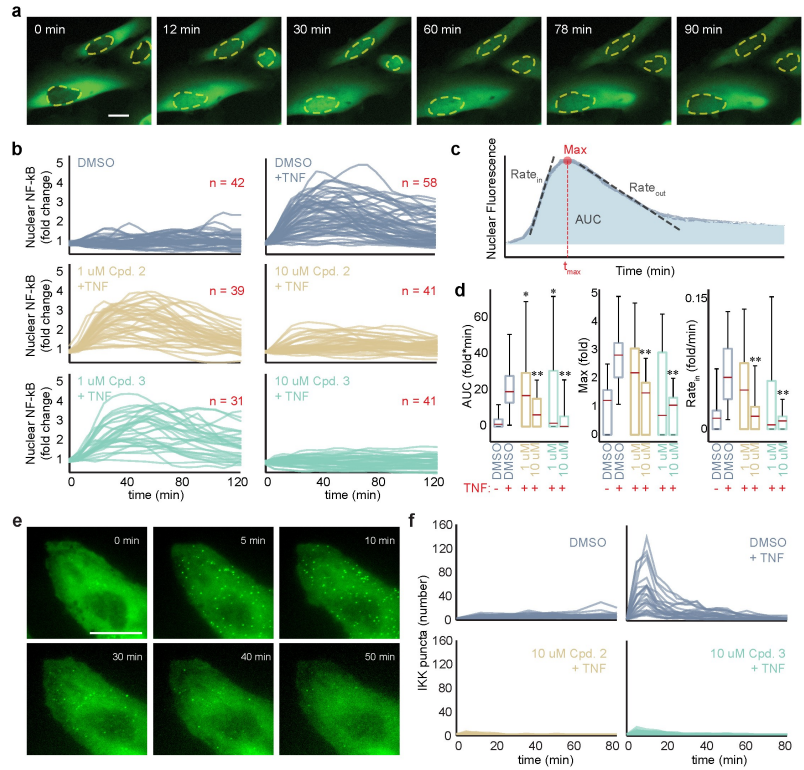


Figure 4.3. Small molecules disruptors of NF- κ B signaling reduce nuclear translocation of NF- κ B and the formation of NEMO puncta in TNF-stimulated cells. (a) Time-lapse images of FP-RelA expressed from its endogenous gene locus in U2OS cells exposed to TNF. The nuclear subcellular compartment is indicated with a broken yellow line. Scale bar 20 μ m for all. **(b)** Single cell time courses of nuclear FP-RelA measure the change in the nuclear abundance of NF- κ B in response to the indicated conditions. Red numbers indicate the number of single cell trajectories in each condition. **(c)** Descriptors used to quantify single cell responses. AUC, Max, and t_{max} , respectively, describe the area under the curve, the maximum, and the time of maximal nuclear FP-RelA fluorescence. Rate_{in} and Rate_{out} describe the maximal rate of nuclear entry and exit. **(d)** Box (first and third quartile) and whisker (1.5 times interquartile range) plots showing the condition-specific variation for descriptors of nuclear FP-RelA localization. Red bars indicate the median; * $p < 0.05$, ** $p < < 10^{-5}$, t test. **(e)** Time-lapse images of FP-IKK expressed from its endogenous gene locus in U2OS cells exposed to TNF. **(f)** Single-cell time courses for the number of FP-IKK puncta in cells stimulated with the indicated conditions. In all TNF conditions, a concentration of 10 ng/mL was used.

4.3 DISCUSSION

Taken together our results show that compounds 2 and 3 inhibit the TNF-induced NF- κ B signaling pathway by limiting the formation of the mature TNFR1 complex. We also highlight the much broader effects of disrupting a pathway component within the larger network, including the downregulation of IKK γ protein expression, and the limitations of single-target molecular modelling as a basis for drug design. The regulatory complexity of the NF- κ B signaling pathway, which enables highly specific and stimulus-dependent transcriptional responses, also confounds drug discovery efforts that do not account for network-scale responses to chemical disruption. Consequently, successful therapeutic intervention in complex signaling pathways may require a network-centric strategy guided explicitly by a compounds' anticipated effects on signaling dynamics as a pharmacologic target. [155]

Correlations in gene expression signatures and single-cell experiments can be used to respectively predict and validate the network effects of bioactive compounds, and structural analysis can further inform on their mechanism of action. Here, our models suggest that compounds 2 and 3 destabilize interactions between TRADD and TRAF-family proteins. Mechanistically, disruption at this upstream junction will preclude ubiquitin scaffold assembly, and this rationalizes the correlations observed between our compounds and knockdowns of UBB and UBC, in addition to other signaling proteins that are recruited to these poly-ubiquitin chains (see Figure 4.1a).

Although the LINCS dataset does not explicitly report the transcriptional response of cells to TNF in the presence of chemical or genetic perturbations, compounds that impinge on TNF-induced dynamics could still be inferred using a machine learning algorithm with prior knowledge of the signaling network. This pipeline therefore represents an alternative strategy to single-target-based drug discovery that can be

more generally applied to discover novel inhibitors of protein subnetworks in a variety of signaling pathways. Because mechanism of action is not constrained *a priori*, it is possible to discover a chemical agent that disrupts multiple points in the same protein subnetwork, or to predict chemical combinations that produce specific network-level responses. It is unlikely that the “magic bullet” drug discovery paradigm will uncover the full therapeutic potential of compounds that modulate dynamic intracellular signals, such as the TNF-induced NF- κ B signaling pathway. Rather, more effective drug development efforts may require approaches like the one presented here, that embrace the complexity of regulatory networks and dynamic phenotypes associated with their disruption.

4.4 METHODS

4.4.1 Analysis of gene expression data

Preparation and analysis of gene expression (GE) data was performed as described in Chapter 3.4. Briefly, gene knockdown (KD) and compound treatment GE signatures were extracted from the LINCS L1000 Phase I and Phase II datasets (GEO accession IDs: GSE70138 and GSE92742). We collected signatures for the 1680 small molecules and 3104 gene KD experiments that had been performed in at least four of the seven most common LINCS cells lines (A549, MCF7, VCAP, HA1E, A375, HCC515, HT19). We hypothesized that compounds that disrupt the TNF-inducible NF- κ B signaling pathway should produce similar network-level effects, and thus similar differential GE signatures, to genetic knockdowns of proteins in the pathway. Thus, for each compound – KD signature pair in our dataset, we computed several cell-specific quantitative features, most importantly:

Direct correlation: the Pearson correlation coefficient between the compound treatment and the gene KD expression signatures in the given cell line, and

Indirect correlation: the fraction of the KD protein's interaction partners, as defined by BioGrid[86], whose respective KD signatures were highly correlated with the compound signature.

Three additional features, quantifying baseline drug activity in the cell and the maximum & average compound-induced differential expression levels of NF- κ B pathway proteins[156], were also calculated and used in subsequent classification.

Using a Random Forest (RF) classifier trained the expression signatures of 152 FDA-approved drugs with known mechanism(s) of action (see Chapter 3.4), features for every compound-KD pair ($n=5,214,720$) were used to predict the probability that the compound would inhibit the KD protein's interaction network. The top-100 predicted interactions for each compound were extracted, and compounds whose predicted targets were enriched in TNF-induced NF- κ B signaling genes ($n=501$) were collected for structural analysis.

4.4.2 Structural analysis

Structural docking of RF – predicted inhibitors proceeded as previously in Chapter 3.4. Briefly, representative crystal structures of TNF-inducible NF- κ B signaling proteins (Supplementary Fig. 1) were mined from the PDB [157], optimizing for sequence coverage, structural resolution, and structural diversity. Domain structures were available for all proteins in Figure 4.1a with the exception of IKK α . Potential small-molecule binding sites on each protein structure were identified by clustering the output of computational solvent mapping software FTMap [119]. RF-predicted inhibitors were docked to predicted binding sites on each protein structure using smina [91], and a prospectively validated pipeline [90, 92]. Three promising candidate inhibitors of TRAF2, which showed both biophysical complementarity

and broad spectrum transcriptomic correlations with knockdowns in the pathway, were purchased from MolPort for experimental validation.

4.4.3 Thermal shift assay and analysis

TRAF2 - compound interactions were measured by fluorescence-based thermal shift using an Applied Biosystems ABI QuantStudio(TM) 6 Flex System. All assay experiments used 1uM GST-TRAF2 (Rockland) per well and 2 X Sybro Orange (Invitrogen) in a buffer containing 50mM HEPES, pH 7.5, 150mM NaCl in a total reaction volume of 15ul in 384 well plates. Compounds were diluted with DMSO, and each reaction had a final DMSO concentration of 1.5%. PCR plates were covered with optical seal, shaken, and centrifuged after protein and compounds were added. The instrument was programmed in the Melt Curve mode and the Standard speed run. The reporter was selected as Rox and None for the quencher. Each melt curve was programmed as follows: 25 °C for 2 min, followed by a 0.05°C increase per second from 25 °C to 99 °C, and finally 99°C for 2min. Fluorescence intensity was collected continuously. In the Melt Curve Filter section, X4 (580 ±10)-M4 (623±14) was selected for the Excitation Filter-Emission Filter. The raw data was extracted in MS-Excel format. Each melt curve was normalized between 0 and 1 and the midpoint of the curve was used to determine the melting temperature.

4.4.4 Establishing EGFP- RELA /IKK γ CRISPR Knock-in Cells

Construction of Repair Templates for EGFP-IKK γ CRISPR Knock-in: The RelA repair template consisted of DNA sequences for a left homology arm (LHA -544bp, chromosome 11_65663376 - chromosome 11_65662383) followed by an EGFP coding sequence with a start codon but no stop codon and a sequence encoding 3x GGSG linker followed by a right homology arm (RHA +557bp, chromosome 11_65662829 - chromosome 11_65662276) were assembled from plasmids synthesized by GeneArt. Synonymous mutations that are not recognized guide RNAs were introduced to prevent interaction the repair template

and Cas9. IKBKG DNA sequences for left homology arm (LHA -861bp, chromosome X 154551142- chromosome X 154552002) and right homology arm (RHA +797bp, chromosome X_154552006 - chromosome X 154552798) were amplified from HeLa genomic DNA using the following primer pairs: IKBKG_LHA_F: 5'GGG CGA ATT GGG CCC GAC GTC GTT TCA CCG TGT TAG CCA GG3', IKBKG_LHA_R: 5' CAC ATC CTT ACC CAG CAG A3'; IKBKG_RHA_F: 5'AGA GTC TCC TCT GGG GAA GC3, IKBKG_RHA_R: 5'CCG CCA TGG CGG CCG GGA GCA TGC GAC GTC AGT CTA GGA AAG AAC TCC CCA GTC3'. In order to generate the fragment containing EGFP overlapping with LHA and RHA, we synthesized the sequence from GeneArt, then we amplified the sequence containing EGFP with the primer pairs: IKBKG_EGFP_F 5' TCT GCT GGG TAA GGA TGT G3', IKBKG_EGFP_R 5' GCT CTT GAT TCT CCT CCA GGC AG 3'. After PCR products were purified, the fragments LHA, RHA, EGFP were cloned to pMK plasmid that was digested with AatII by gibson assembly from NEB.

Construction of Guide RNA: The guide RNAs were designed by the CRISPR Design Tool (<http://crispr.mit.edu>). Oligonucleotide pairs Rel A sg1 (top): 5'-CACCGCTCGTCTGTAGTGCACGCCG-3', Rel A sg1 (bottom): 5'-AAACCGGCGTGCACTACAGACGAGC-3'; RELA Sg2 (top) 5'-CACCGAGAGGCGGAAATGCGCCGCC-3', RELA Sg2 (bottom) 5'- AAACCGGCGCGCATTTCGCTCTC-3'; IKBKG Sg1 (top) 5'-CACCGGCAGCAGATCAGGACGTAC-3', IKBKG Sg1 (bottom) 5'-AAACGTACGTCTGATCTGCTGCC-3'; and IKBKG Sg2 (top) 5'-CACCGCTGCACCATCTCACACAGT-3', IKBKG Sg2 (bottom) 5'-AAACTGTGTGAGATGGTGCAGC-3' were cloned into the vector pSpCas9n (BB)-2A-Puro (PX462) (Addgene). The pSpCas9n (BB)-2A-Puro-IKKγ_gRNAs vector encoded the guide RNA and the Cas9 nuclease with D10A nickase mutant.

Transfection and Clone Isolation: U2OS cells (2×10^5 cells per well) were seeded in 6-well plates in complete growth medium. The following day, with pSpCas9n (BB)-2A-Puro-RELA/IKKγ_gRNAs and repair

template donor plasmids were linearized using BGLII, and cells were transfected using FuGENE HD (Promega) with a transfection reagent to DNA ratio of 3.5 to 1 and a total DNA amount of 4 μ g. After two weeks, cells were subjected to single cell sorting into 96-well plates using Beckman Coulter MoFlo Astrios High Speed. Cells underwent clonal isolation and a positive clone was identified via western blot and confirmed by live-cell imaging.

4.4.5 Western blot analysis

U2OS cells (parental and expressing EGFP-RelA/IKK γ via CRISPR Knock-in) were cultured for 24 hrs in complete growth medium. After treatments, cells were lysed in SDS-based lysis buffer consisting of 120 mM Tris-Cl, pH 6.8, 4% SDS supplemented with protease and phosphatase inhibitors at 4°C for 30 min. Protein extracts were clarified by centrifugation at 4°C at 12,000 $\times g$ for 10 min. Lysate protein levels were quantified by BCA assay (Pierce). Samples were separated by SDS-PAGE, 25 μ g total protein per lane, then transferred to PVDF membranes. Blocking was done in 5% milk in TBS for 1 hour. Primary antibodies directed at RelA and β -actin (#4764 and #3700 respectively; Cell Signaling Technology), IKK γ and GAPDH (sc-8330 and sc25778 respectively; Santa Cruz) were diluted in 5% milk in TBS-T and incubated overnight at 4°C. Alexa 680/800-conjugated secondary antibodies (LICOR) were used in combination with an Odyssey (LI-COR) scanner for detection and quantification of band intensities.

4.4.6 Live-cell imaging and analysis

Live cells were imaged in an environmentally controlled chamber (37°C, 5% CO₂) on a DeltaVision Elite microscope equipped with a pco.edge sCMOS camera and an Insight solid-state illumination module (GE). U2OS cells expressing FP-RelA/IKK γ were seeded at a density of 25000 cells/well 24 hours prior to live-cell imaging experiments on no. 1.5 glass bottom 96 well imaging plates (Matriplate). For imaging of FP-RelA nuclear translocation, live-cells were pre-treated with DMSO or indicated concentrations of compounds

for 2 hours before exposure to 100ng/ml TNF. Wide-field epifluorescence and DIC images were collected using a 20x LUCPLFLN objective (0.45NA; Olympus). Cells were imaged for at least 30 minutes prior to addition of compounds. For detection of IKK γ puncta, live-cells were pre-treated with DMSO or indicated concentration of compounds for 2 hours before exposure to 100ng/ml TNF. Wide-field epifluorescence and DIC images were collected using a 60x LUCPLFLN objective. For all treatments, cytokine mixtures were prepared and pre-warmed so that addition of 120 μ L added to wells results in a final concentration as indicated. Time-lapse images were collected over at least 4 fields per condition with a temporal resolution of 5 minutes per frame. Quantification of nuclear FP-RelA localization and the formation IKK γ puncta from flat-field and background corrected images was performed using customized scripts in Matlab and ImageJ.

4.4.7 Fixed-cell immunofluorescence and analysis

For fixed-cell measurement of endogenous RelA (Supplementary Figure 4), U2OS cells were seeded into plastic bottom 96 well imaging plates (Fisher) at 6000 cell/well 24 hours prior to treatment. On the day of the experiment, media containing TNF was prepared at 15X the desired concentration for each well. Timing of TNF treatment was planned so fixation (0, 10, 30, 60, 90, 120 minutes) occurred simultaneously for all time points at the same time. Pre-warmed 15X cytokine mixture was spiked into wells and mixed. Between treatments the cells remained in environmentally controlled conditions (37°C and 5% CO₂).

At time zero, media was removed from the wells, 185 μ L of PBS was used to wash the wells, and wells were incubated at room temp in 120 μ L of 4% paraformaldehyde (PFA) in 1X PBS for 10 minutes. Wells were then washed 3X three minutes with 185 μ L 1X PBS and then incubated in 120 μ L 100% methanol for 10 min at room temp. Next wells were washed 3X three minutes in PBS-T (1XPBS 0.1% Tween 20) followed by 120 μ L of primary antibody solution (3% BSA PBS-T, 1 μ g/mL NF- κ B p65 F-6 (sc-8008; Santa Cruz)). Plates were wrapped in para-film and left to incubate at 4°C overnight. The following morning, wells were

washed 3X five minutes in 185 μ L PBS-T followed by incubation for 1 hour in 120 μ L of the secondary antibody solution (3% BSA PBS-T, 4 μ g/mL Goat anti-Mouse IgG Alexa Fluor 647 (Thermo Fisher)). 185 μ L PBS-T was used to wash the wells for 5 minutes and they were put into 120ul Hoechst solution (PBS-T, 200ng/mL Hoechst) for 20 min. Finally, wells were washed five minutes with PBST and then 185 μ L PBS was used to fill the wells and keep the cells hydrated during imaging. Cells were imaged using Delta Vision Elite imaging system at 20x magnification with a LUCPLFN objective (0.45NA; Olympus). Analysis was done using Cell Profiler to segment cells and quantify median nuclear intensity values. Further analysis was performed using custom scripts in MATLAB.

4.5 SUPPLEMENTARY FIGURES

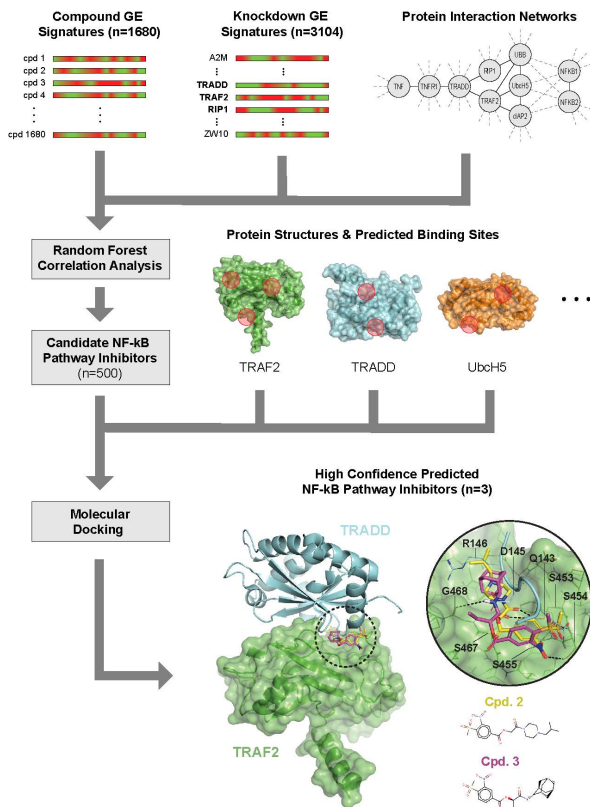


Figure 4.4. Prediction pipeline used to identify small molecule inhibitors of TNF-inducible NF-kB signaling. Input includes cell-specific gene expression (GE) signatures from 1680 bioactive small molecules and 3104 gene knockdowns, taken from the LINCS L1000 dataset [147], and the protein interaction networks of knockdown genes, inferred from their BioGrid [86] – defined interaction partners. Correlations between compound and knockdown GE signatures and their distribution on the TNF-inducible NF-kB pathway (see Figure 4.1a) are evaluated by a random forest classifier to predict candidate inhibitors. Structural models of pathway proteins are mined from the PDB [157] and used as molecular docking targets for candidates. Docking results are assessed to identify high-confidence predicted inhibitors.

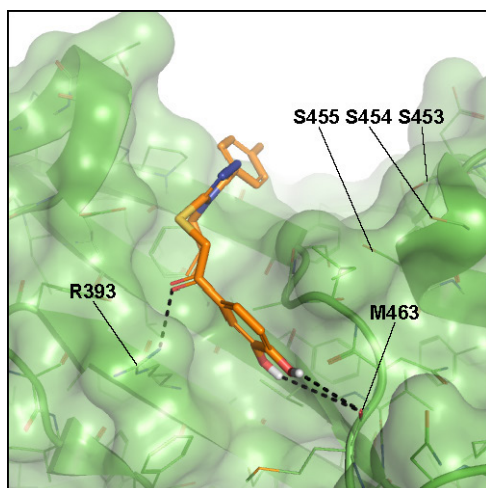


Figure 4.5. Predicted binding mode of compound 1 to TRADD-binding interface of TRAF2. Hydrogen bonds are indicated with dotted lines.

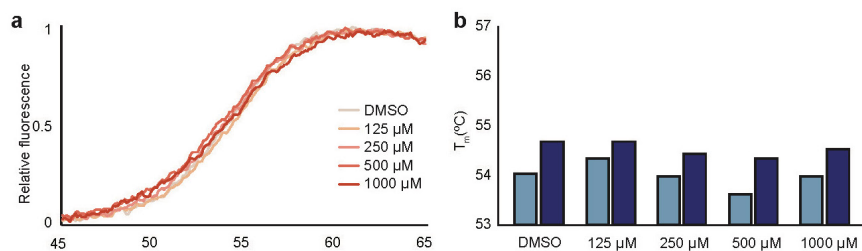


Figure 4.6. Thermal shift assays indicate no clear effect of compound 1 on TRAF2 stability. (a) Normalized melt curve of full length TRAF2 in the presence of DMSO or indicated concentrations of compound 1. (b) Melting temperature of TRAF2 in the presence of compound 1 are not significantly altered in replicate experiments.

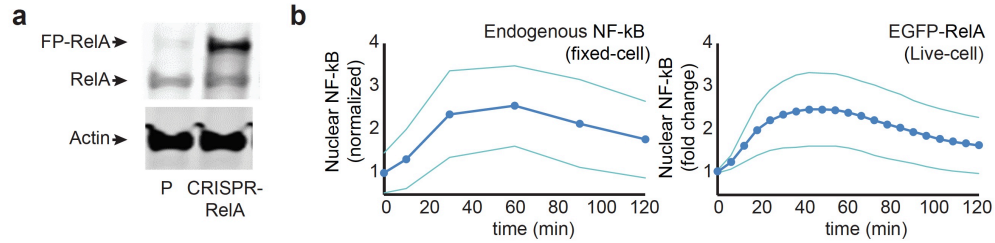


Figure 4.7. Quantification of FP-RelA expression in U2OS cells. (a) Western blot of RelA in lysates from parental U2OS cells (P) and U2OS cells that were modified using CRISPR to express EGFP-RelA. The molecular weight of the dominant FP-RelA band in the CRISPR-modified cell line is shifted upward by 32 kDa, consistent with the expected molecular weight of the EGFP fusion protein. The presence of the wild type RelA band in the CRISPR-modified cell line suggests that only one allele of the RelA-encoding gene integrated the EGFP-encoding sequence. (b) Subcellular localization of RelA from fixed-cell immunofluorescence images of parental U2OS cells (left) and FP-RelA quantified from CRISPR-modified live cells (right) exposed to 10 ng/mL TNF show similar temporal dynamics.

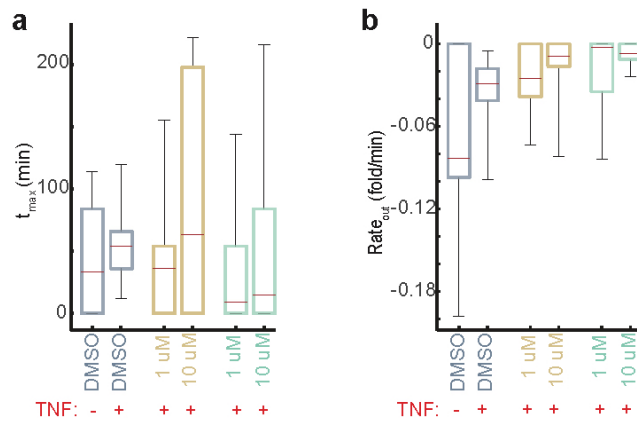


Figure 4.8. Other descriptors of nuclear FP-RelA. Descriptors t_{max} (a) and $Rate_{out}$ (b) do not show statistically significant differences in response to TNF.

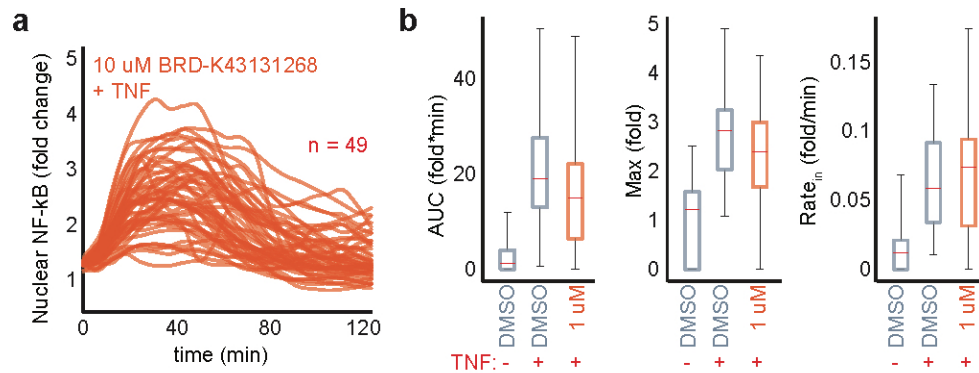


Figure 4.9. Compound 1 does not have a significant effect on FP-RelA translocation. (a) Single cell time courses measure the change in nuclear abundance of FP-RelA in cells exposed to 10ng/mL TNF after pre-incubation with compound 1. **(b)** Descriptors of nuclear FP-RelA dynamics do not change significantly even in the presence of a high concentration of compound 1.

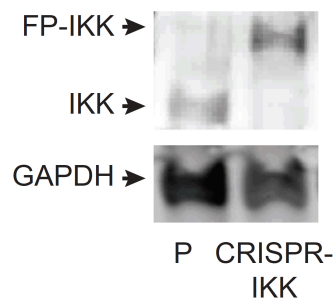


Figure 4.10. Western blot of IKK γ . Western blot of IKK γ in lysates from parental U2OS cells (P) and U2OS cells that were modified using CRISPR to express EGFP-IKK γ . The molecular weight of the FP-IKK γ band in the CRISPR-modified cell line is shifted upward by 32 kDa, consistent with the expected molecular weight of the EGFP fusion protein. The absence of wild type IKK γ in the CRISPR-modified cell line suggests that both alleles of the IKK γ -encoding gene integrated the EGFP sequence.

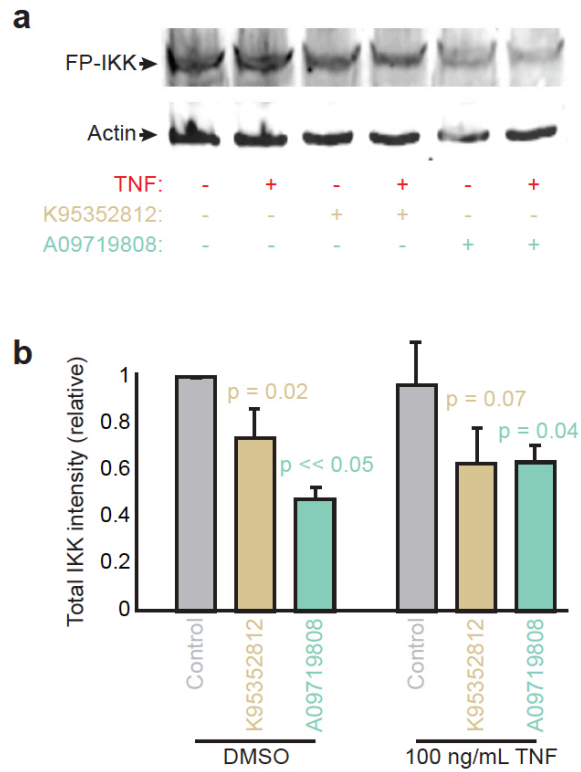


Figure 4.11. IKK γ expression in the presence of compounds 2 and 3. (a) Western blot of IKK γ in lysates from CRISPR-modified U2OS cells in the indicated conditions. (b) Quantification of Actin-corrected IKK band intensity, normalized to control cells that were not pre-treated with compounds, suggest that the presence of compounds 2 and 3 significantly downregulate the expression of IKK γ . Indicated p values were calculated from t tests of biological triplicates.

5.0 DRUGQUERY: STRUCTURE-BASED SMALL MOLECULE VIRTUAL SCREENING OF HUMAN PROTEINS

5.1 INTRODUCTION

Rising efficacy- and safety-related failure rates in pharmaceutical development programs [3-5] suggest the decline of the target-centric “magic bullet” drug discovery paradigm and its emphasis on single-target *in-vitro* potency [158]. In its place are emerging more global, network-centric approaches that embrace the complex regulatory environments in which drugs function and seek to account for multi-target effects [159-161]. Considering that the average drug currently on the market exhibits activity towards six molecular targets [162, 163], often resulting in unexpected and unwanted side effects [164-167], this shift in focus away from single proteins and towards multi-target networks is perhaps a natural step. The new drug discovery paradigm of polypharmacology seeks to anticipate and exploit the off-target effects of promiscuous small molecules [159, 160, 168-170], raising an immediate need for robust methods of predicting the protein targets of bioactive compounds, or target fishing.

As experimental approaches to target fishing are often cost- and time-inefficient [171], more efficient computational approaches that leverage existing data have become popular [172, 173]. Many are ligand- or network-based methods that extrapolate from known interactions of compounds that are structurally or functionally similar to the query compound [65-69, 174-176]. The main alternatives are structure-based methods, which evaluate the 3-D complementarity of potential ligand-target pairs using their atomic structures [70-72, 111, 177]. With the rapid growth of protein crystal structures in the PDB [157],

structure-based methods like reverse molecular docking offer a number of advantages over ligand-based similarity searching such as being unbiased with respect to novel chemical scaffolds and being able to predict the query compound's binding mode [72]. These features are highly desirable in a drug discovery context, where new chemistries are often being tested and where knowledge of binding mode is necessary for lead optimization [178].

Despite these advantages, there currently exist no user friendly, publicly available tools for reverse docking at a large scale (1000+ potential targets). Published reverse-docking software has either been decommissioned [70], is prohibitively slow [71], or is unavailable for public use [177, 179]. To fill this void, we present DrugQuery (DQ): a fast, user-friendly, and publicly available web server for reverse-docking-based target fishing. The DQ target library currently contains 7957 predicted binding sites on 2069 high quality crystal structures of 1245 unique human proteins, all easily searchable and available for download. User-submitted small molecules, accepted in all standard molecular file formats, are docked against the DQ library using the smina implementation of AutoDock Vina [91, 180]. Results, including a ranked list of potential targets and the query compound's predicted binding modes, are returned in hours for most small molecules, facilitating fast and easy target fishing for any chemist or biologist.

On a validation set of 95 FDA-approved drugs with known protein targets, DQ achieved top-10 and top-100 target prediction accuracies of 35% and 58% percent, respectively, with at least one known target ranked in to top decile for 68% of compounds. Within the same validation set, DQ correctly identified multiple known protein targets for 10 drugs with promiscuous activity. Remarkably, without re-docking, 76% of successful predictions were associated with binding modes < 2 Angstrom RMSD from known bound configurations. On a separate validation set of 102 congeneric FXR-binding compounds taken from the

2017 D3R Grand Challenge 2, DQ achieved top-10 and top-100 target prediction accuracies of 27% and 72%, respectively, with FXR ranked in to top decile for 86% of compounds.

5.2 METHODS

5.2.1 The DrugQuery target library

The DQ target library currently contains 2069 high quality crystal structures of 1245 human genes. Representative structures were extracted from the Protein Data Bank (PDB) [157] using a previously described greedy algorithm⁸ that considers all available human protein structures of a gene and selects a small number that maximize total sequence coverage and structural resolution (see Chapter 3.4.8). The selection process accounts for conformational diversity by allowing multiple representative structures of a single domain if they are separated by backbone RMSD above a sliding threshold, which iteratively increases if too many representative structures are identified. The selection process does not explicitly consider the presence of cocrystallized ligands in protein structures, which reduces DQ's bias towards known druggable targets. The distribution of representative structure counts for targets in the DQ library is shown in Figure 5.1a.

⁸ https://github.com/npabon/generate_gene_models

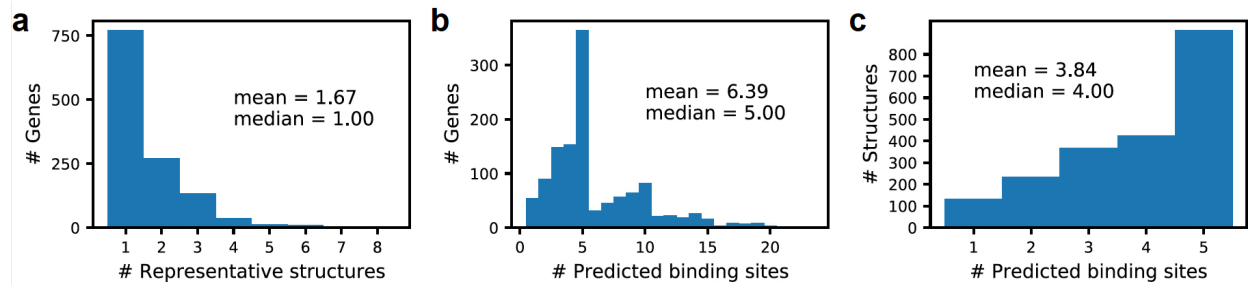


Figure 5.1. Profiles of protein structures and predicted binding sites in the DrugQuery library. (a) Distribution of the representative structure counts for the 1245 genes in the DQ library. (b) Distribution of predicted binding site counts for DQ genes. (c) Distribution of predicted binding site counts for DQ protein target structures.

Druggable binding sites on representative target structures are identified using the Atlas computational solvent mapping software employed by the FTMap web server [119], which has been demonstrated to identify known ligand-binding pockets with high accuracy [181]. Atlas output is clustered to combine fragments separated by a minimum distance of less than five Angstroms and (up to) the five largest clusters for each protein structure are stored in the DQ library, though occasionally fewer than five clusters are produced for a given target. The distributions of predicted binding site counts per-gene and per-target structure are shown in Figure 5.1b,c.

5.2.2 Docking and scoring

Docking to predicted binding sites is performed using the smina⁹ implementation of Autodock Vina [91, 180]. Smina achieved state-of-the-art accuracy in both pose prediction and affinity ranking in recent community wide challenges including Drug Design Data Resource (D3R) and Community Structure-Activity Resource (CSAR) [90, 91, 182]. Docking to predicted binding sites is performed using a box padding of four

⁹ <https://sourceforge.net/projects/smina/>

Angstroms and standard rigid-receptor smina sampling and exhaustiveness parameters. For each compound – target – binding-site combination up to nine score-ranked docked poses are generated and saved to the DQ database. When dockings are completed for a given query compound, target structures are ranked by the top docking score achieved across each target's predicted binding sites.

5.2.3 RMSD analysis of predicted binding modes

During validation, when a crystal structure of a validation compound's known binding pose was available in the PDB, we compared it to the binding mode predicted by DQ. We computed root-mean-squared deviations (RMSDs) between heavy atoms in the predicted vs. known poses using OpenBabel [183]. For compounds without reference binding poses available in the PDB, we searched for crystal structures of similar compounds bound to the same protein target. Structural similarity between compounds was evaluated in terms of the Tanimoto coefficient, computed using OpenBabel. "Similar" compounds were defined as having Tanimoto similarity > 0.7 to the validation compound. When a similar compound was available, the Python RDKit¹⁰ module was used to identify the maximum common scaffold (MCS) between it and the validation compound. We then used MCS heavy atoms to compute a partial RMSD between the DQ-predicted poses of the validation compound and the known binding mode of the similar compound.

5.2.4 Web server

The DQ interface was written using the Python web framework Django (2.0.4). DQ compounds, targets, and docking results are stored in a MySQL database (14.14). All cheminformatic operations, including validating user uploads and computing compound similarities, are carried out using the Open Babel (2.3.1) Python wrapper PyBel [183]. 3Dmol [184] is used to create interactive, three dimensional renderings of user-uploaded compounds and the targets in the DQ library. Job runtimes scale roughly with the number

¹⁰ <http://www.rdkit.org/>

of rotatable bonds in the compound and can range from one hour for small, rigid ligands to several hours for large, flexible molecules.

5.3 USING DRUGQUERY

5.3.1 Uploading a new small molecule

On the DQ upload page (Figure 5.2a) users submit a ligand file and enter an email address to which a job completion notification will be sent. DQ accepts all standard small molecule file formats recognized by Open Babel [183] (SMI, SDF, PDB, MOL2, CIF, etc...). Before docking, DQ screens uploaded compounds against a database of previously uploaded compounds to check for duplicates. If the compound already exists in the DQ database, docking is skipped and the user is redirected to its existing results page. If not, DQ adds the compound to the database, assigns it a unique numerical ID reflecting the chronology of its upload, submits a docking job to the queue, and initializes an empty results page. The user is then redirected to the newly created results page (Figure 5.2b), which is populated upon completion of the docking job.

DrugQuery [About](#) [Dock](#) [Compounds](#) [Genes](#) [Queue](#)

a

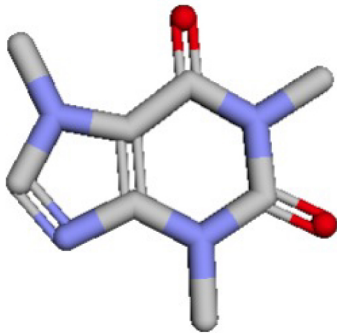
Upload a compound to DrugQuery
 Accepts all molecular file formats recognized by [Open Babel](#). Only the first compound in the input file will be docked.

Upload file: No file chosen

Email:

DrugQuery [About](#) [Dock](#) [Compounds](#) [Genes](#) [Queue](#)

b



Compound 1

Smiles:
Cn1cnc2c1c(=O)n(C)c(=O)n2C

Current docking status:
 Queued

Top 100 Predicted Targets
 Target predictions will become available once dockings for this compound are complete.

Figure 5.2. The DrugQuery user interface. (a) The DQ compound upload page. (b) The unpopulated results page for a newly-submitted compound.

5.3.2 Tracking a job

Tracking the status of a docking job is achieved using by visiting the associated compound's results page, accessible at *drugquery.csb.pitt.edu/compounds/<id>/*, which displays the status of the docking job (Figure 5.2b). To check a queued job's place in line, users can inspect the DQ queue, accessible at *drugquery.csb.pitt.edu/queue/*. Once a docking job is completed, a link to the query compound's results page is sent to the email address associated with the compound upload.

5.3.3 Downloading results

As shown in Figure 5.2b, a compound's results page displays an interactive 3Dmol[184] rendering of the ligand and a table listing of the top-100 predicted protein targets. Each table entry lists the HUGO gene symbol of the target, the PDB and chain IDs of the protein model, the pocket ID of the predicted binding site, and the smina-predicted affinity score. There are several download options available to the user for retrieving results: (1) a columnated text file containing the complete target rankings and affinity scores, (2) a zipped directory containing the compound docking models (SDF format) and target structures (PDB format) for the top-100 predicted targets, or (3) a zipped directory containing the complete set of compound docking models and structures for all DQ targets. Docking results (2) and (3) are organized hierarchically in directories by gene name, PDB ID, and chain ID.

5.4 VALIDATION

5.4.1 Predicting targets of FDA-approved drugs

To demonstrate the usage of DQ we constructed a validation set of 95 FDA approved drugs with known targets listed in DrugBank [185] and ChEMBL [186] whose target genes were present in the DQ library

(Additional File 5.1). Hereby referred to as Validation Set 1 (VS1), the distributions of known target counts, masses, and rotatable bonds for these compounds are shown in Figure 5.3. As several compounds had multiple known targets (Figure 5.3A), VS1 contained total of 169 pairwise interactions.

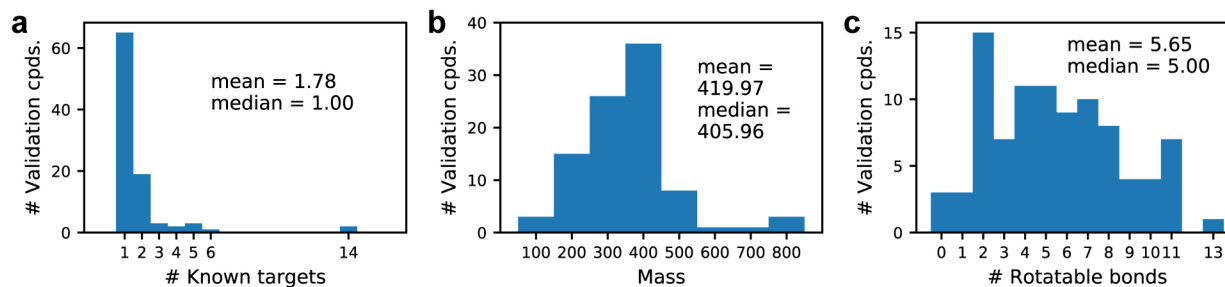


Figure 5.3. Properties of FDA-approved compounds in Validation Set 1 (n=95). (a) Distribution of the number of known target genes present in the DQ library. (b) Molecular weight distribution. (c) Distribution of number of rotatable bonds.

For each VS1 compound, DQ ranked the 2069 potential target structures by maximum docking score across the targets' predicted binding sites. For 33 (35%), 55 (58%), and 65 (68%) of VS1 compounds, respectively, DQ predicted at least one known target in the top-10, top-100, and top-10% of potential targets. For 36 (21%), 66 (39%), and 84 (50%) of VS1 interactions, respectively, DQ predicted the known target in the top-10, top-100, and top-10%. Out of 30 compounds in VS1 with multiple known targets in the DQ library, DQ correctly ranked multiple targets in the top-10% for 10 (33%). The compound- and interaction-specific receiver operating characteristic (ROC) curves produced by DQ on VS1 are shown in Figure 5.4 and have area-under-the-curves (AUCs) of 0.89 and 0.80, respectively, demonstrating accuracies comparable to recent ligand-based [187] and genomic [80, 81] target prediction methods of a similar scale.

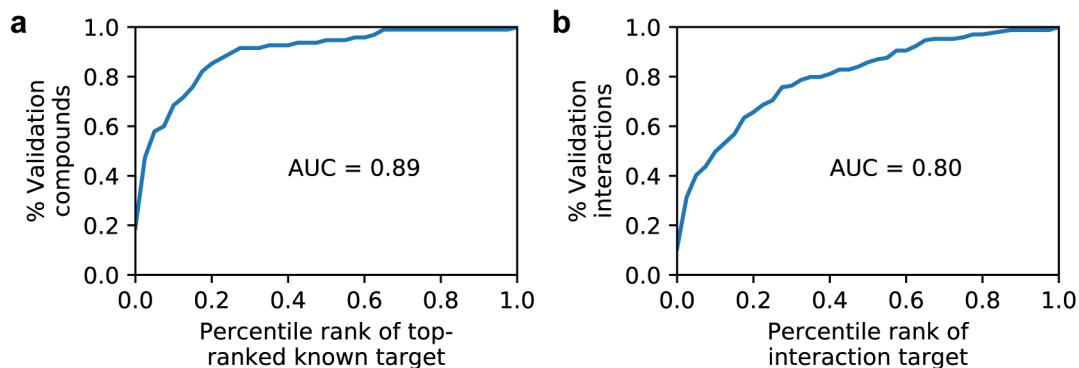


Figure 5.4. DrugQuery target prediction accuracy on Validation Set 1. (a) Compound-specific ROC curve, generated by sorting the 95 VS1 compounds by the rank of each one's top-ranked known target. (b) Interaction-specific ROC curve, generated by sorting the 169 VS1 interactions by the rank of the known target for the relevant compound.

One important advantage of DQ over alternative target prediction methods, however, is its ability to predict the 3D binding modes of the query compound to potential targets. 54 of the 84 VS1 interactions for which DQ predicted the known target in the top-10% had corresponding cocrystals in the PDB depicting the known binding mode of the interaction compound or a similar compound sharing a common scaffold (see Chapter 5.2.3 – RMSD analysis of predicted binding modes). Remarkably, for 34 (63%) of these interactions, the first DQ-predicted pose had RMSD < 2 Angstroms from the native pose, whereas if we considered the top-5 predicted poses then 41 (76%) had RMSD < 2 Angstroms (Figure 5.5). It is worth noting that re-docking – docking a compound into the receptor structure from a cocrystal in which it is already bound to the compound being docked – was explicitly avoided during validation by removing the offending target structures in the DQ library from consideration.

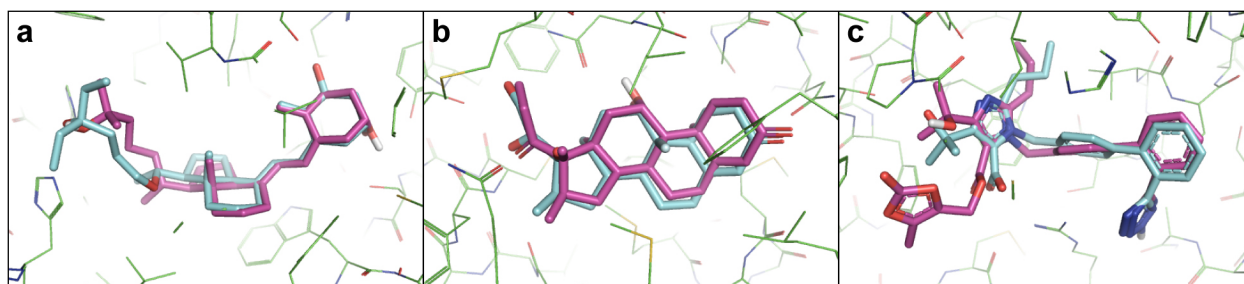


Figure 5.5. DrugQuery predicts native-like poses for compounds in Validation Set 1. (a) The DQ-predicted binding pose of calcifediol (magenta) aligned to a crystal structure (PDB ID: 1IE8) of KH1 (cyan) bound to the VDR protein (green). (b) The DQ-predicted binding pose of rimexolone (magenta) aligned to a crystal structure (PDB ID: 3MNE) of DEX (cyan) bound to the NR3C1 protein (green). (c) The DQ-predicted binding pose of olmesartan (magenta) aligned to a crystal structure (PDB ID: 4ZUD) of OLM (cyan) bound to the AGTR1 protein (green).

5.4.2 Predicting targets of non-drug bioactive compounds

To evaluate the performance of DQ in predicting the targets of non-drug small molecules, we constructed a second validation set (VS2) of 102 congeneric compounds from the 2017 community-wide D3R Grand Challenge 2. All VS2 compounds had known activity against the human protein NR1H4 (FXR), for which there are two representative structures (PDB IDs 4OIV & 3OKI) in the DQ library. The mass and rotatable bond distributions of VS2 compounds are shown in Figure 5.6a,b. For 28 (27%), 73 (72%), and 88 (86%) of VS2 compounds, respectively, DQ predicted FXR in the top-10, top-100, and top 10% of 2069 ranked potential targets. DQ's VS2 ROC curve is shown in Figure 5.6c and has an AUC of 0.94.

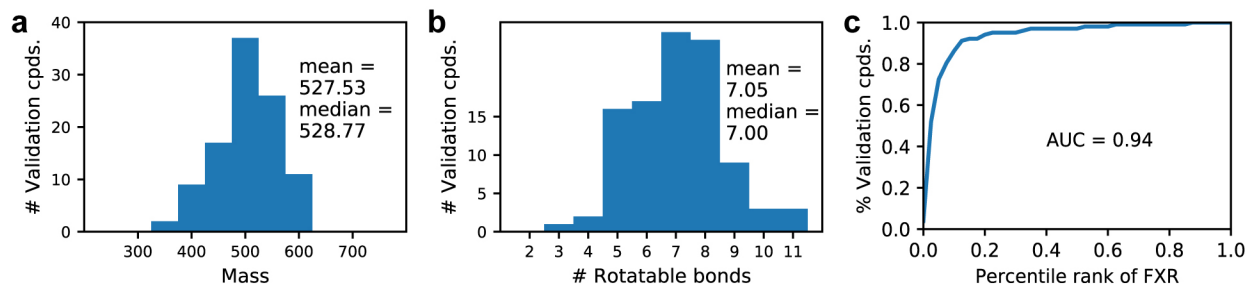


Figure 5.6. Properties of compounds in Validation Set 2 (n=102) and DrugQuery target prediction accuracy. (a) Molecular weight distribution. (b) Distribution of number of rotatable bonds. (c) Compound-specific ROC curve, generated by sorting the 102 VS2 compounds by FXR's position among their ranked potential targets.

As part of the D3R challenge, bound poses for 36 VS2 compounds were released to the PDB and we used them to evaluate the compounds' DQ-predicted poses. For 10 (28%) of these compounds, the first DQ-predicted pose had RMSD < 2 Angstroms from the native pose, whereas if we considered the top-5 predicted poses then 16 (44%) had RMSD < 2 Angstroms (Figure 5.7). We observed that DQ's comparatively weaker pose prediction accuracy for VS2 vs. VS1 was due largely to the fact that several VS2 FXR structures deviated significantly from the representative FXR structures in the DQ library (Figure 5.8a). Predicted poses for these VS2 compounds were thus often "flipped" with respect to their true binding modes - maintaining specific, native-like hydrogen bond contacts but "swapping" the positions of nonspecific hydrophobic groups (Figure 5.8b,c).

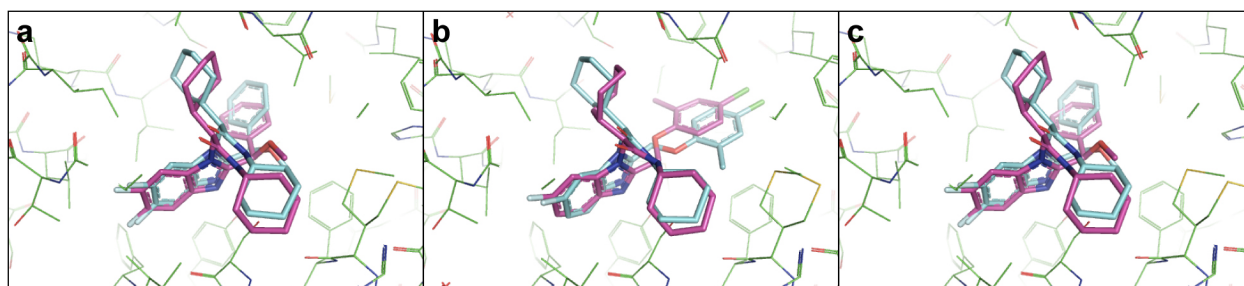


Figure 5.7. Native-like predicted poses for compounds in Validation Set 2. DQ-predicted poses (magenta) aligned to bound cocrystal structures (green & cyan) for (a) FXR19, (b) FXR20, and (c) FXR22.

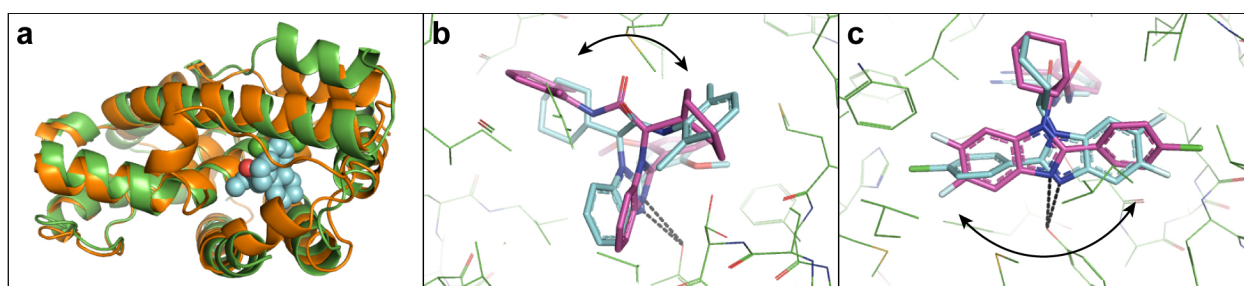


Figure 5.8. Unique receptor conformations in Validation Set 2 produce “swapped” hydrophobic contacts in predicted poses. (a) Bound crystal structure of FXR4 (green & cyan) aligned to DQ’s representative FXR structure (orange, PDB ID 3OKI), highlighting significant conformational differences in binding-pocket-adjacent protein regions. (b) DQ-predicted pose (magenta) of FXR14 aligned to bound cocrystal structure (green & cyan). Dashed black lines denote conserved, native-like hydrogen bonds. Solid black arrow indicates the “swapped” nonspecific hydrophobic groups of FXR14. (c) DQ-predicted pose (magenta) of FXR25 aligned to bound cocrystal structure (green & cyan). Native contacts and hydrophobic “swapping” indicated as in (b).

5.5 SUMMARY

Predicting the protein targets of bioactive compounds has many critical applications in drug discovery including side effect prediction, drug repurposing, and the design of multi-target drugs. In lieu of costly wet lab target identification pipelines, ligand- and structure-based computational approaches have emerged. Of these, docking-based multi-target screening has the significant advantage of predicting the structural binding modes of query compounds. With the growth of available 3D protein structures in the PDB, docking-based methods are becoming increasingly well suited for target prediction and have in several cases predicted novel interactions that were subsequently verified experimentally [179, 188]. Currently, however, there are no fast and easy-to-use tools available to the scientific community for multi-target docking at a large scale.

Presented here is a public web server called DrugQuery (DQ) that meets this need – providing a simple and intuitive interface for to predict potential protein targets for small molecules of interest. The DQ library currently contains 2069 domain structures of potential protein targets from 1245 unique human genes and has the potential to scale up to the 10,000+ unique human proteins in the PDB. Using DQ, users can upload a compound of interest in any standard molecular file format and in a short time, a few hours for most small molecules, can easily download the complete DQ potential target rankings and the predicted binding modes of the query compound.

DQ has been extensively tested for both target identification and binding site prediction on a chemically diverse validation set of 95 multi-target FDA-approved drugs, as well as a second validation set of 102 congeneric FXR-binding compounds from the 2017 D3R Grand Challenge 2. In both cases, target ranking ROCs achieve AUCs above 0.89 (Figure 5.4, Figure 5.6c), demonstrating significant discriminative power in league with recent, non-structure-based, computational target prediction methods [80, 81, 187].

Furthermore, DQ demonstrated multi-target prediction accuracy by correctly identifying multiple known protein targets for 10/30 multi-target drugs in VS1. In binding pose prediction, DQ achieved impressive accuracy considering that redocking was explicitly avoided. In VS1, DQ predicted poses with < 2 Angstrom RMSD from the native pose for 76% of compounds with bound cocrystal structures. In VS2, sub-2 Angstrom RMSDs were achieved for 44% of compounds.

DQ is in active development and there are a number of avenues for future improvement. In addition to expanding the DQ target library, we plan to implement compound-specific score normalization strategies to account for the so called “reverse docking scoring bias” - a phenomena that has been characterized for most available docking software [189] in which target predictions are artificially biased towards structures with larger, more hydrophobic cavities. Potential improvements to the representative structure selection process are also being considered, such as explicitly considering the presence of bound ligands and attempting to maximize binding-site diversity.

5.6 ADDITIONAL FILES

Additional File 5.1. (additional_file_5.1.xlsx) Table of compounds in Validation Set 1 and their known protein targets that have representative structures in the DurgQuery target library.

6.0 CONCLUSIONS AND FUTURE RESEARCH

The research presented in this dissertation has focused on improving rational drug discovery by applying insights into the biophysics that regulate protein interactions and the network-scale effects of disrupting them. We focused on four major challenges in this arena: protein flexibility and selective promiscuity, protein target prediction for bioactive small molecules, disrupting complex signaling networks, and large-scale structure-based target screening. Below, we will briefly recap our contributions to each of these areas and the computational approaches employed therein.

In Chapter 2.0 we discussed our efforts to model how structural flexibility at the surface of immune checkpoint receptor PD-1 facilitates ‘selectively promiscuous’ binding – binding specifically to multiple protein ligands with structurally distinct binding interfaces. Using molecular dynamics simulations, we identified evolutionarily conserved “trigger” motifs on the ligands’ interfaces that, upon recognition by PD-1, displace polar Asn66 and transforms PD-1’s interface from flat and hydrophilic to flexible and hydrophobic. Ligand-specific, trigger-adjacent interactions then stabilize distinct bound-like receptor interfaces: a flat hydrophobic patch for PD-L1 and a large hydrophobic cavity for PD-L2. The importance of trigger interactions was demonstrated by a recent crystal structure of the blockbuster PD-1 – targeting antibody pembrolizumab, which, although having evolved via an entirely different pathway than PD-1’s cognate ligands, exploits analogous triggering interactions to displace Asn66. This, and the structural modelling of a recently patented PD-1 inhibiting macrocycle suggest the potential for triggering to guide rational drug design against challenging, high-impact targets like PD-1. Future work in this area will

consist of: (1) incorporating conserved trigger motifs in virtual screening for inhibitors potentially capable of transforming hard-to-drug interfaces like PD-1's into surfaces more amenable to small molecule binding, and (2) evaluating the generality of the trigger model of selective promiscuity by studying other flexible, multi-ligand regulatory proteins.

In Chapter 3.0 we examined a hybrid genomic & structural pipeline for small molecule protein target prediction. Based on the hypothesis that small molecule protein inhibitors should yield similar transcriptomic responses in live cells to the genetic knockdown of the target protein, we trained a random forest classifier to predict drug-target interactions from gene expression data. We found that in addition to *direct* correlations between the gene expression signatures of drugs and the knockdowns of potential targets, accurate predictions required considering *indirect* correlations between the drug and other knockdowns in the potential target's pathway. By refining our genomic predictions with structure-based modelling, we demonstrate that we can accurately predict the known targets of FDA-approved drugs as well as predicting previously unknown interactions, of which we validate several. Future work in this area will involve searching for ways to improve random forest classification accuracy, such as distinguishing between inhibitors and agonists during training or testing additional gene expression-based features. We may also explore the feasibility of employing alternative classification models like neural networks.

In Chapter 4.0 we describe an extension of the genomic insight we gained from Chapter 3.0 and its application to identifying small molecule disruptors of the TNF-induced NF- κ B signaling pathway. By virtually screening for compounds that produced broad-spectrum transcriptomic correlations with the genetic inhibition of proteins in the TNFR1 signaling complex, we identified two compounds that inhibited IKK recruited and prevented NF- κ B in live cell fluorescence assays. Structural modelling suggested a

mechanism of action involving disruption of the native TRADD-TRAF2 interface, which would likely have downstream effects affecting ubiquitin scaffolding and recruitment of downstream signaling proteins.

Finally, in Chapter 5.0 we present DrugQuery a first-of-its-kind public web server for docking-based small molecule target prediction. With over 2000 target structures and 1200 unique human genes currently represented in the DrugQuery library, we validated the server extensively in both target ranking and binding pose prediction contexts. DrugQuery's simple web interface was designed with special consideration for potential users with limited computational cheminformatic or structural modelling experience, such that any biologist or chemist with a computer can quickly and easily obtain a ranked target list and predicted binding modes for their small molecule of interest. As a project still in active development, we plan to scale-up the DrugQuery target library to exploit the full diversity of the Protein Data Bank as well as evaluate alternative strategies for rational representative structure selection. We may also make changes to DrugQuery's backend docking and scoring pipeline as the Camacho group continues to participate in and learn from community-wide challenges in small molecule affinity ranking and pose prediction.

BIBLIOGRAPHY

1. Mavromoustakos T, Durdagi S, Koukoulitsa C, Simcic M, Papadopoulos MG, Hodoscek M, et al. Strategies in the rational drug design. *Curr Med Chem*. 2011;18(17):2517-30. Epub 2011/05/17. PubMed PMID: 21568895.
2. Schneider G, Fechner U. Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov*. 2005;4(8):649-63. Epub 2005/08/02. doi: 10.1038/nrd1799. PubMed PMID: 16056391.
3. Waring MJ, Arrowsmith J, Leach AR, Leeson PD, Mandrell S, Owen RM, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat Rev Drug Discov*. 2015;14(7):475-86. Epub 2015/06/20. doi: 10.1038/nrd4609. PubMed PMID: 26091267.
4. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov*. 2004;3(8):711-5. Epub 2004/08/03. doi: 10.1038/nrd1470. PubMed PMID: 15286737.
5. Scannell JW, Bosley J. When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. *PLoS One*. 2016;11(2):e0147215. Epub 2016/02/11. doi: 10.1371/journal.pone.0147215. PubMed PMID: 26863229; PubMed Central PMCID: PMC4749240.
6. Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. Structure-based virtual screening for drug discovery: a problem-centric review. *The AAPS journal*. 2012;14(1):133-41. Epub 2012/01/28. doi: 10.1208/s12248-012-9322-0. PubMed PMID: 22281989; PubMed Central PMCID: PMC3282008.
7. Dugger SA, Platt A, Goldstein DB. Drug development in the era of precision medicine. *Nat Rev Drug Discov*. 2018;17(3):183-96. Epub 2017/12/09. doi: 10.1038/nrd.2017.226. PubMed PMID: 29217837.
8. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol*. 1999;293(2):321-31. doi: 10.1006/jmbi.1999.3110. PubMed PMID: 10550212.
9. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform*. 2000;11:161-71. PubMed PMID: 11700597.
10. Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, et al. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol*. 2006;2(8):e100. doi: 10.1371/journal.pcbi.0020100. PubMed PMID: 16884331; PubMed Central PMCID: PMC1526461.

11. Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, et al. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A*. 2006;103(22):8390-5. doi: 10.1073/pnas.0507916103. PubMed PMID: 16717195; PubMed Central PMCID: PMC1482503.
12. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*. 2004;337(3):635-45. doi: 10.1016/j.jmb.2004.02.002. PubMed PMID: 15019783.
13. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, et al. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J Proteome Res*. 2007;6(5):1917-32. doi: 10.1021/pr060394e. PubMed PMID: 17391016; PubMed Central PMCID: PMC1482503.
14. Liu J, Faeder JR, Camacho CJ. Toward a quantitative theory of intrinsically disordered proteins and their function. *Proc Natl Acad Sci U S A*. 2009;106(47):19819-23. doi: 10.1073/pnas.0907710106. PubMed PMID: 19903882; PubMed Central PMCID: PMC2775701.
15. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol*. 2002;323(3):573-84. PubMed PMID: 12381310.
16. Schon O, Friedler A, Bycroft M, Freund SM, Fersht AR. Molecular mechanism of the interaction between MDM2 and p53. *J Mol Biol*. 2002;323(3):491-501. PubMed PMID: 12381304.
17. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. *J Mol Biol*. 1999;285(5):2177-98. PubMed PMID: 9925793.
18. Betts MJ, Sternberg MJ. An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng*. 1999;12(4):271-83. PubMed PMID: 10325397.
19. Cheng X, Veverka V, Radhakrishnan A, Waters LC, Muskett FW, Morgan SH, et al. Structure and interactions of the human programmed cell death 1 receptor. *J Biol Chem*. 2013;288(17):11771-85. Epub 2013/02/19. doi: 10.1074/jbc.M112.448126. PubMed PMID: 23417675; PubMed Central PMCID: PMC3636866.
20. Zak KM, Kitel R, Przetocka S, Golik P, Guzik K, Musielak B, et al. Structure of the Complex of Human Programmed Death 1, PD-1, and Its Ligand PD-L1. *Structure*. 2015;23(12):2341-8. doi: 10.1016/j.str.2015.09.010. PubMed PMID: 26602187.
21. Lazar-Molnar E, Yan Q, Cao E, Ramagopal U, Nathenson SG, Almo SC. Crystal structure of the complex between programmed death-1 (PD-1) and its ligand PD-L2. *Proc Natl Acad Sci U S A*. 2008;105(30):10483-8. Epub 2008/07/22. doi: 10.1073/pnas.0804453105. PubMed PMID: 18641123; PubMed Central PMCID: PMC2492495.
22. Lin DY, Tanaka Y, Iwasaki M, Gittis AG, Su HP, Mikami B, et al. The PD-1/PD-L1 complex resembles the antigen-binding Fv domains of antibodies and T cell receptors. *Proc Natl Acad Sci U S A*. 2008;105(8):3011-6. Epub 2008/02/22. doi: 10.1073/pnas.0712278105. PubMed PMID: 18287011; PubMed Central PMCID: PMC2268576.

23. Ma B, Kumar S, Tsai CJ, Nussinov R. Folding funnels and binding mechanisms. *Protein Eng.* 1999;12(9):713-20. Epub 1999/10/03. PubMed PMID: 10506280.
24. Tsai CJ, Kumar S, Ma B, Nussinov R. Folding funnels, binding funnels, and protein function. *Protein Sci.* 1999;8(6):1181-90. Epub 1999/07/01. doi: 10.1110/ps.8.6.1181. PubMed PMID: 10386868; PubMed Central PMCID: PMC2144348.
25. Koshland DE. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc Natl Acad Sci U S A.* 1958;44(2):98-104. Epub 1958/02/01. PubMed PMID: 16590179; PubMed Central PMCID: PMC335371.
26. Boehr DD, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol.* 2009;5(11):789-96. doi: 10.1038/nchembio.232. PubMed PMID: 19841628; PubMed Central PMCID: PMCPMC2916928.
27. Hoang J, Prosser RS. Conformational selection and functional dynamics of calmodulin: a (19)F nuclear magnetic resonance study. *Biochemistry.* 2014;53(36):5727-36. doi: 10.1021/bi500679c. PubMed PMID: 25148136.
28. Hammes GG, Chang YC, Oas TG. Conformational selection or induced fit: a flux description of reaction mechanism. *Proc Natl Acad Sci U S A.* 2009;106(33):13737-41. doi: 10.1073/pnas.0907195106. PubMed PMID: 19666553; PubMed Central PMCID: PMCPMC2728963.
29. Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein-protein interactions. *Proc Natl Acad Sci U S A.* 2004;101(31):11287-92. Epub 2004/07/23. doi: 10.1073/pnas.0401942101. PubMed PMID: 15269345; PubMed Central PMCID: PMC509196.
30. Domling A, Holak TA. Programmed death-1: therapeutic success after more than 100 years of cancer immunotherapy. *Angew Chem Int Ed Engl.* 2014;53(9):2286-8. Epub 2014/01/30. doi: 10.1002/anie.201307906. PubMed PMID: 24474145.
31. Couzin-Frankel J. Breakthrough of the year 2013. Cancer immunotherapy. *Science.* 2013;342(6165):1432-3. doi: 10.1126/science.342.6165.1432. PubMed PMID: 24357284.
32. Zarganes-Tzitzikas T, Konstantinidou M, Gao Y, Krzemien D, Zak K, Dubin G, et al. Inhibitors of programmed cell death 1 (PD-1): a patent review (2010-2015). *Expert Opin Ther Pat.* 2016;26(9):973-7. doi: 10.1080/13543776.2016.1206527. PubMed PMID: 27367741.
33. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. *Protein Sci.* 1996;5(12):2438-52. doi: 10.1002/pro.5560051206. PubMed PMID: 8976552; PubMed Central PMCID: PMCPMC2143314.
34. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* 1998;7(9):1884-97. doi: 10.1002/pro.5560070905. PubMed PMID: 9761470; PubMed Central PMCID: PMCPMC2144175.
35. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, et al. Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol.* 2007;25(1):71-5. doi: 10.1038/nbt1273. PubMed PMID: 17211405.

36. Cozzini P, Kellogg GE, Spyraakis F, Abraham DJ, Costantino G, Emerson A, et al. Target flexibility: an emerging consideration in drug discovery and design. *J Med Chem.* 2008;51(20):6237-55. Epub 2008/09/13. doi: 10.1021/jm800562d. PubMed PMID: 18785728; PubMed Central PMCID: PMC2701403.
37. DeLisi C. The biophysics of ligand-receptor interactions. *Q Rev Biophys.* 1980;13(2):201-30. PubMed PMID: 7015404.
38. Northrup SH, Erickson HP. Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proc Natl Acad Sci U S A.* 1992;89(8):3338-42. PubMed PMID: 1565624; PubMed Central PMCID: PMC48862.
39. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res.* 2004;32(Web Server issue):W96-9. doi: 10.1093/nar/gkh354. PubMed PMID: 15215358; PubMed Central PMCID: PMC441492.
40. Horita S, Nomura Y, Sato Y, Shimamura T, Iwata S, Nomura N. High-resolution crystal structure of the therapeutic antibody pembrolizumab bound to the human PD-1. *Sci Rep.* 2016;6:35297. doi: 10.1038/srep35297. PubMed PMID: 27734966; PubMed Central PMCID: PMC5062252.
41. Lee JY, Lee HT, Shin W, Chae J, Choi J, Kim SH, et al. Structural basis of checkpoint blockade by monoclonal antibodies in cancer immunotherapy. *Nat Commun.* 2016;7:13354. doi: 10.1038/ncomms13354. PubMed PMID: 27796306; PubMed Central PMCID: PMC5095608.
42. Na Z, Yeo SP, Bharath SR, Bowler MW, Balikci E, Wang CI, et al. Structural basis for blocking PD-1-mediated immune suppression by therapeutic antibody pembrolizumab. *Cell Res.* 2017;27(1):147-50. doi: 10.1038/cr.2016.77. PubMed PMID: 27325296; PubMed Central PMCID: PMC5223238.
43. Champ PC, Camacho CJ. FastContact: a free energy scoring tool for protein-protein complex structures. *Nucleic Acids Res.* 2007;35(Web Server issue):W556-60. Epub 2007/06/01. doi: 10.1093/nar/gkm326. PubMed PMID: 17537824; PubMed Central PMCID: PMC1933237.
44. Miller MM, Mapelli C, Allen MP, Bowsher MS, Boy KM, Gillis EP, et al. Macrocyclic inhibitors of the pd-1/pd-l1 and cd80(b7-1)/pd-l1 protein/protein interactions. *Google Patents*; 2014.
45. Vainio MJ, Johnson MS. Generating conformer ensembles using a multiobjective genetic algorithm. *J Chem Inf Model.* 2007;47(6):2462-74. doi: 10.1021/ci6005646. PubMed PMID: 17892278.
46. Pawson T, Scott JD. Signaling through scaffold, anchoring, and adaptor proteins. *Science.* 1997;278(5346):2075-80. PubMed PMID: 9405336.
47. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph.* 1996;14(1):33-8, 27-8. Epub 1996/02/01. PubMed PMID: 8744570.
48. D.A. Case VB, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker,

- J. Wang, R.M. Wolf, X. Wu and P.A. Kollman. AMBER 14. University of California, San Francisco; 2014.
49. Gotz AW, Williamson MJ, Xu D, Poole D, Le Grand S, Walker RC. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J Chem Theory Comput.* 2012;8(5):1542-55. Epub 2012/05/15. doi: 10.1021/ct200909j. PubMed PMID: 22582031; PubMed Central PMCID: PMC3348677.
 50. The PyMOL Molecular Graphics System. 1.5.0.1 ed: Schrödinger, LLC; 2010.
 51. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics.* 2009;10:168. Epub 2009/06/03. doi: 10.1186/1471-2105-10-168. PubMed PMID: 19486540; PubMed Central PMCID: PMC2700099.
 52. Schmidtke P, Le Guilloux V, Maupetit J, Tuffery P. fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.* 2010;38(Web Server issue):W582-9. Epub 2010/05/19. doi: 10.1093/nar/gkq383. PubMed PMID: 20478829; PubMed Central PMCID: PMC2896101.
 53. Edwards AM, Isserlin R, Bader GD, Frye SV, Willson TM, Yu FH. Too many roads not taken. *Nature.* 2011;470(7333):163-5. doi: 10.1038/470163a. PubMed PMID: 21307913.
 54. Grueneberg DA, Degot S, Pearlberg J, Li W, Davies JE, Baldwin A, et al. Kinase requirements in human cells: I. Comparing kinase requirements across various cell types. *Proc Natl Acad Sci U S A.* 2008;105(43):16472-7. doi: 10.1073/pnas.0808019105. PubMed PMID: 18948591; PubMed Central PMCID: PMCPMC2575444.
 55. Fedorov O, Muller S, Knapp S. The (un)targeted cancer kinome. *Nat Chem Biol.* 2010;6(3):166-9. doi: 10.1038/nchembio.297. PubMed PMID: 20154661.
 56. Swinney DC, Anthony J. How were new medicines discovered? *Nat Rev Drug Discov.* 2011;10(7):507-19. doi: 10.1038/nrd3480. PubMed PMID: 21701501.
 57. Bleicher KH, Bohm HJ, Muller K, Alanine AI. Hit and lead generation: beyond high-throughput screening. *Nat Rev Drug Discov.* 2003;2(5):369-78. doi: 10.1038/nrd1086. PubMed PMID: 12750740.
 58. Pritchard JF, Jurima-Romet M, Reimer ML, Mortimer E, Rolfe B, Cayen MN. Making better drugs: Decision gates in non-clinical drug development. *Nat Rev Drug Discov.* 2003;2(7):542-53. doi: 10.1038/nrd1131. PubMed PMID: 12815380.
 59. Mayr LM, Bojanic D. Novel trends in high-throughput screening. *Curr Opin Pharmacol.* 2009;9(5):580-8. doi: 10.1016/j.coph.2009.08.004. PubMed PMID: 19775937.
 60. Persidis A. High-throughput screening. Advances in robotics and miniturization continue to accelerate drug lead identification. *Nat Biotechnol.* 1998;16(5):488-9. doi: 10.1038/nbt0598-488. PubMed PMID: 9592401.

61. Gregori-Puigjane E, Setola V, Hert J, Crews BA, Irwin JJ, Lounkine E, et al. Identifying mechanism-of-action targets for drugs and probes. *Proc Natl Acad Sci U S A*. 2012;109(28):11178-83. doi: 10.1073/pnas.1204524109. PubMed PMID: 22711801; PubMed Central PMCID: PMC3396511.
62. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? *Nat Rev Drug Discov*. 2006;5(12):993-6. doi: 10.1038/nrd2199. PubMed PMID: 17139284.
63. Drews J. Drug discovery: a historical perspective. *Science*. 2000;287(5460):1960-4. PubMed PMID: 10720314.
64. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. *Nature*. 2009;462(7270):175-81. doi: 10.1038/nature08506. PubMed PMID: 19881490; PubMed Central PMCID: PMC2784146.
65. Yamanishi Y, Kotera M, Kanehisa M, Goto S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*. 2010;26(12):i246-54. doi: 10.1093/bioinformatics/btq176. PubMed PMID: 20529913; PubMed Central PMCID: PMC2881361.
66. Martinez-Jimenez F, Marti-Renom MA. Ligand-target prediction by structural network biology using nAnnoLyze. *PLoS Comput Biol*. 2015;11(3):e1004157. doi: 10.1371/journal.pcbi.1004157. PubMed PMID: 25816344; PubMed Central PMCID: PMC4376866.
67. Nickel J, Gohlke BO, Erehman J, Banerjee P, Rong WW, Goede A, et al. SuperPred: update on drug classification and target prediction. *Nucleic Acids Res*. 2014;42(Web Server issue):W26-31. doi: 10.1093/nar/gku477. PubMed PMID: 24878925; PubMed Central PMCID: PMC4086135.
68. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Res*. 2014;42(Web Server issue):W32-8. doi: 10.1093/nar/gku293. PubMed PMID: 24792161; PubMed Central PMCID: PMC4086140.
69. Lo YC, Senese S, Li CM, Hu Q, Huang Y, Damoiseaux R, et al. Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. *PLoS Comput Biol*. 2015;11(3):e1004153. doi: 10.1371/journal.pcbi.1004153. PubMed PMID: 25826798; PubMed Central PMCID: PMC4380459.
70. Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, et al. TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res*. 2006;34(Web Server issue):W219-24. doi: 10.1093/nar/gkl114. PubMed PMID: 16844997; PubMed Central PMCID: PMC1538869.
71. Wang JC, Chu PY, Chen CM, Lin JH. idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. *Nucleic Acids Res*. 2012;40(Web Server issue):W393-9. doi: 10.1093/nar/gks496. PubMed PMID: 22649057; PubMed Central PMCID: PMC3394295.
72. Rognan D. Structure-Based Approaches to Target Fishing and Ligand Profiling. *Mol Inform*. 2010;29(3):176-87. doi: 10.1002/minf.200900081. PubMed PMID: WOS:000276336900002.

73. Meslamani J, Li J, Sutter J, Stevens A, Bertrand HO, Rognan D. Protein-ligand-based pharmacophores: generation and utility assessment in computational ligand profiling. *J Chem Inf Model.* 2012;52(4):943-55. doi: 10.1021/ci300083r. PubMed PMID: 22480372.
74. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, et al. Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.* 2008;36(Database issue):D866-70. doi: 10.1093/nar/gkm815. PubMed PMID: 17932051; PubMed Central PMCID: PMCPMC2238822.
75. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science.* 2006;313(5795):1929-35. doi: 10.1126/science.1132939. PubMed PMID: 17008526.
76. Lamb J. The Connectivity Map: a new tool for biomedical research. *Nat Rev Cancer.* 2007;7(1):54-60. doi: 10.1038/nrc2044. PubMed PMID: 17186018.
77. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaekar P, Ferriero R, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci U S A.* 2010;107(33):14621-6. doi: 10.1073/pnas.1000138107. PubMed PMID: 20679242; PubMed Central PMCID: PMCPMC2930479.
78. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med.* 1998;4(11):1293-301. doi: 10.1038/3282. PubMed PMID: 9809554.
79. Cosgrove EJ, Zhou Y, Gardner TS, Kolaczyk ED. Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia. *Bioinformatics.* 2008;24(21):2482-90. doi: 10.1093/bioinformatics/btn476. PubMed PMID: 18779235; PubMed Central PMCID: PMCPMC2732281.
80. Isik Z, Baldow C, Cannistraci CV, Schroeder M. Drug target prioritization by perturbed gene expression and network information. *Sci Rep.* 2015;5:17417. doi: 10.1038/srep17417. PubMed PMID: 26615774; PubMed Central PMCID: PMCPMC4663505.
81. Laenen G, Thorrez L, Bornigen D, Moreau Y. Finding the targets of a drug by integration of gene expression data with a protein interaction network. *Mol Biosyst.* 2013;9(7):1676-85. doi: 10.1039/c3mb25438k. PubMed PMID: 23443074.
82. Andy Liaw MW. Classification and regression by randomforest. *R news.* 2002;2(3):18-22.
83. Qi Y, Bar-Joseph Z, Klein-Seetharaman J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins.* 2006;63(3):490-500. doi: 10.1002/prot.20865. PubMed PMID: 16450363; PubMed Central PMCID: PMCPMC3250929.
84. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, et al. Functional discovery via a compendium of expression profiles. *Cell.* 2000;102(1):109-26. PubMed PMID: 10929718.
85. Schenone M, Dancik V, Wagner BK, Clemons PA. Target identification and mechanism of action in chemical biology and drug discovery. *Nat Chem Biol.* 2013;9(4):232-40. doi: 10.1038/nchembio.1199. PubMed PMID: 23508189; PubMed Central PMCID: PMCPMC5543995.

86. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* 2015;43(Database issue):D470-8. doi: 10.1093/nar/gku1204. PubMed PMID: 25428363; PubMed Central PMCID: PMC4383984.
87. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, et al. Cell type-specific gene expression differences in complex tissues. *Nat Methods.* 2010;7(4):287-9. doi: 10.1038/nmeth.1439. PubMed PMID: 20208531; PubMed Central PMCID: PMC3699332.
88. Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics.* 2006;7:3. doi: 10.1186/1471-2105-7-3. PubMed PMID: 16398926; PubMed Central PMCID: PMC1363357.
89. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD, Rodgers JR, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Arch Biochem Biophys.* 1978;185(2):584-91. Epub 1978/01/30. PubMed PMID: 626512.
90. Ye Z, Baumgartner MP, Wingert BM, Camacho CJ. Optimal strategies for virtual screening of induced-fit and flexible target in the 2015 D3R Grand Challenge. *J Comput Aided Mol Des.* 2016;30(9):695-706. doi: 10.1007/s10822-016-9941-0. PubMed PMID: 27573981; PubMed Central PMCID: PMC5079819.
91. Koes DR, Baumgartner MP, Camacho CJ. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J Chem Inf Model.* 2013;53(8):1893-904. doi: 10.1021/ci300604z. PubMed PMID: 23379370; PubMed Central PMCID: PMC3726561.
92. Baumgartner MP, Camacho CJ. Choosing the Optimal Rigid Receptor for Docking and Scoring in the CSAR 2013/2014 Experiment. *J Chem Inf Model.* 2016;56(6):1004-12. doi: 10.1021/acs.jcim.5b00338. PubMed PMID: 26222931; PubMed Central PMCID: PMC4744803.
93. Koes DR, Pabon NA, Deng X, Phillips MA, Camacho CJ. A Teach-Discover-Treat Application of ZincPharmer: An Online Interactive Pharmacophore Modeling and Virtual Screening Tool. *PLoS One.* 2015;10(8):e0134697. doi: 10.1371/journal.pone.0134697. PubMed PMID: 26258606; PubMed Central PMCID: PMC4530941.
94. Der CJ, Krontiris TG, Cooper GM. Transforming genes of human bladder and lung carcinoma cell lines are homologous to the ras genes of Harvey and Kirsten sarcoma viruses. *Proc Natl Acad Sci U S A.* 1982;79(11):3637-40. PubMed PMID: 6285355; PubMed Central PMCID: PMC346478.
95. Schubbert S, Shannon K, Bollag G. Hyperactive Ras in developmental disorders and cancer. *Nat Rev Cancer.* 2007;7(4):295-308. doi: 10.1038/nrc2109. PubMed PMID: 17384584.
96. Maurer T, Garrenton LS, Oh A, Pitts K, Anderson DJ, Skelton NJ, et al. Small-molecule ligands bind to a distinct pocket in Ras and inhibit SOS-mediated nucleotide exchange activity. *Proc Natl Acad Sci U S A.* 2012;109(14):5299-304. doi: 10.1073/pnas.1116510109. PubMed PMID: 22431598; PubMed Central PMCID: PMC3325706.
97. Welsch ME, Kaplan A, Chambers JM, Stokes ME, Bos PH, Zask A, et al. Multivalent Small-Molecule Pan-RAS Inhibitors. *Cell.* 2017;168(5):878-89 e29. doi: 10.1016/j.cell.2017.02.006. PubMed PMID: 28235199; PubMed Central PMCID: PMC5362268.

98. Ostrem JM, Shokat KM. Direct small-molecule inhibitors of KRAS: from structural insights to mechanism-based design. *Nat Rev Drug Discov.* 2016;15(11):771-85. doi: 10.1038/nrd.2016.139. PubMed PMID: 27469033.
99. Fetis SK, Guterres H, Kearney BM, Buhrman G, Ma B, Nussinov R, et al. Allosteric effects of the oncogenic RasQ61L mutant on Raf-RBD. *Structure.* 2015;23(3):505-16. doi: 10.1016/j.str.2014.12.017. PubMed PMID: 25684575.
100. Bueno M, Temiz NA, Camacho CJ. Novel modulation factor quantifies the role of water molecules in protein interactions. *Proteins.* 2010;78(15):3226-34. doi: 10.1002/prot.22805. PubMed PMID: 20665475.
101. Paul I, Ghosh MK. A CHIPotle in physiology and disease. *Int J Biochem Cell Biol.* 2015;58:37-52. doi: 10.1016/j.biocel.2014.10.027. PubMed PMID: 25448416.
102. Meacham GC, Patterson C, Zhang W, Younger JM, Cyr DM. The Hsc70 co-chaperone CHIP targets immature CFTR for proteasomal degradation. *Nat Cell Biol.* 2001;3(1):100-5. doi: 10.1038/35050509. PubMed PMID: 11146634.
103. Zhang M, Windheim M, Roe SM, Pegg M, Cohen P, Prodromou C, et al. Chaperoned ubiquitylation--crystal structures of the CHIP U box E3 ubiquitin ligase and a CHIP-Ubc13-Uev1a complex. *Mol Cell.* 2005;20(4):525-38. doi: 10.1016/j.molcel.2005.09.023. PubMed PMID: 16307917.
104. Irwin JJ, Shoichet BK. ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model.* 2005;45(1):177-82. Epub 2005/01/26. doi: 10.1021/ci049714+. PubMed PMID: 15667143; PubMed Central PMCID: PMC1360656.
105. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 2006;34(Database issue):D668-72. doi: 10.1093/nar/gkj067. PubMed PMID: 16381955; PubMed Central PMCID: PMC1347430.
106. Vanhaesebroeck B, Alessi DR. The PI3K-PDK1 connection: more than just a road to PKB. *Biochem J.* 2000;346 Pt 3:561-76. PubMed PMID: 10698680; PubMed Central PMCID: PMC1220886.
107. Gao X, Harris TK. Role of the PH domain in regulating in vitro autophosphorylation events required for reconstitution of PDK1 catalytic activity. *Bioorg Chem.* 2006;34(4):200-23. doi: 10.1016/j.bioorg.2006.05.002. PubMed PMID: 16780920.
108. Masters TA, Calleja V, Armoogum DA, Marsh RJ, Applebee CJ, Laguerre M, et al. Regulation of 3-phosphoinositide-dependent protein kinase 1 activity by homodimerization in live cells. *Sci Signal.* 2010;3(145):ra78. doi: 10.1126/scisignal.2000738. PubMed PMID: 20978239.
109. Komander D, Fairservice A, Deak M, Kular GS, Prescott AR, Peter Downes C, et al. Structural insights into the regulation of PDK1 by phosphoinositides and inositol phosphates. *EMBO J.* 2004;23(20):3918-28. doi: 10.1038/sj.emboj.7600379. PubMed PMID: 15457207; PubMed Central PMCID: PMC1347430.

110. Wang L, P. Wipf, and X.-Q. Xie. HTDocking- identifying possible targets for small molecules by high throughput docking algorithm. 2012.
111. Liu X, Ouyang S, Yu B, Liu Y, Huang K, Gong J, et al. PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Res.* 2010;38(Web Server issue):W609-14. doi: 10.1093/nar/gkq300. PubMed PMID: 20430828; PubMed Central PMCID: PMCPMC2896160.
112. Koes DR, Camacho CJ. ZINCPharmer: pharmacophore search of the ZINC database. *Nucleic Acids Res.* 2012;40(Web Server issue):W409-14. Epub 2012/05/04. doi: 10.1093/nar/gks378. PubMed PMID: 22553363; PubMed Central PMCID: PMC3394271.
113. Dobson CM. Chemical space and biology. *Nature.* 2004;432(7019):824-8. doi: 10.1038/nature03192. PubMed PMID: 15602547.
114. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 2012;40(Database issue):D1100-7. doi: 10.1093/nar/gkr777. PubMed PMID: 21948594; PubMed Central PMCID: PMCPMC3245175.
115. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database--2009 update. *Nucleic Acids Res.* 2009;37(Database issue):D767-72. doi: 10.1093/nar/gkn892. PubMed PMID: 18988627; PubMed Central PMCID: PMCPMC2686490.
116. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 2004;32(Database issue):D258-61. doi: 10.1093/nar/gkh036. PubMed PMID: 14681407; PubMed Central PMCID: PMCPMC308770.
117. Navlakha S, He X, Faloutsos C, Bar-Joseph Z. Topological properties of robust biological and computational networks. *J R Soc Interface.* 2014;11(96):20140283. doi: 10.1098/rsif.2014.0283. PubMed PMID: 24789562; PubMed Central PMCID: PMCPMC4032542.
118. Bakan A, Meireles LM, Bahar I. ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics.* 2011;27(11):1575-7. doi: 10.1093/bioinformatics/btr168. PubMed PMID: 21471012; PubMed Central PMCID: PMCPMC3102222.
119. Kozakov D, Grove LE, Hall DR, Bohnuud T, Mottarella SE, Luo L, et al. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat Protoc.* 2015;10(5):733-55. Epub 2015/04/10. doi: 10.1038/nprot.2015.043. PubMed PMID: 25855957.
120. Dharmiah S, Bindu L, Tran TH, Gillette WK, Frank PH, Ghirlando R, et al. Structural basis of recognition of farnesylated and methylated KRAS4b by PDEdelta. *Proc Natl Acad Sci U S A.* 2016;113(44):E6766-E75. doi: 10.1073/pnas.1615316113. PubMed PMID: 27791178; PubMed Central PMCID: PMCPMC5098621.
121. Berndsen CE, Wolberger C. A spectrophotometric assay for conjugation of ubiquitin and ubiquitin-like proteins. *Anal Biochem.* 2011;418(1):102-10. doi: 10.1016/j.ab.2011.06.034. PubMed PMID: 21771579; PubMed Central PMCID: PMCPMC3178097.

122. Zhang H, Amick J, Chakravarti R, Santarriaga S, Schlanger S, McGlone C, et al. A bipartite interaction between Hsp70 and CHIP regulates ubiquitination of chaperoned client proteins. *Structure*. 2015;23(3):472-82. doi: 10.1016/j.str.2015.01.003. PubMed PMID: 25684577; PubMed Central PMCID: PMC4351142.
123. Sheffield P, Garrard S, Derewenda Z. Overcoming expression and purification problems of RhoGDI using a family of "parallel" expression vectors. *Protein Expr Purif*. 1999;15(1):34-9. doi: 10.1006/prep.1998.1003. PubMed PMID: 10024467.
124. Todi SV, Scaglione KM, Blount JR, Basrur V, Conlon KP, Pastore A, et al. Activity and cellular functions of the deubiquitinating enzyme and polyglutamine disease protein ataxin-3 are regulated by ubiquitination at lysine 117. *J Biol Chem*. 2010;285(50):39303-13. doi: 10.1074/jbc.M110.181610. PubMed PMID: 20943656; PubMed Central PMCID: PMC2998082.
125. Faggiano S, Menon RP, Kelly GP, McCormick J, Todi SV, Scaglione KM, et al. Enzymatic production of mono-ubiquitinated proteins for structural studies: The example of the Josephin domain of ataxin-3. *FEBS Open Bio*. 2013;3:453-8. doi: 10.1016/j.fob.2013.10.005. PubMed PMID: 24251111; PubMed Central PMCID: PMC3829987.
126. Assimon VA, Southworth DR, Gestwicki JE. Specific Binding of Tetratricopeptide Repeat Proteins to Heat Shock Protein 70 (Hsp70) and Heat Shock Protein 90 (Hsp90) Is Regulated by Affinity and Phosphorylation. *Biochemistry*. 2015;54(48):7120-31. doi: 10.1021/acs.biochem.5b00801. PubMed PMID: 26565746.
127. Dettori R, Sonzogni S, Meyer L, Lopez-Garcia LA, Morrice NA, Zeuzem S, et al. Regulation of the interaction between protein kinase C-related protein kinase 2 (PRK2) and its upstream kinase, 3-phosphoinositide-dependent protein kinase 1 (PDK1). *J Biol Chem*. 2009;284(44):30318-27. doi: 10.1074/jbc.M109.051151. PubMed PMID: 19723632; PubMed Central PMCID: PMC2781587.
128. Zhang H, Neimanis S, Lopez-Garcia LA, Arencibia JM, Amon S, Stroba A, et al. Molecular mechanism of regulation of the atypical protein kinase C by N-terminal domains and an allosteric small compound. *Chem Biol*. 2014;21(6):754-65. doi: 10.1016/j.chembiol.2014.04.007. PubMed PMID: 24836908.
129. Schulze JO, Saladino G, Busschots K, Neimanis S, Suss E, Odadzic D, et al. Bidirectional Allosteric Communication between the ATP-Binding Site and the Regulatory PIF Pocket in PDK1 Protein Kinase. *Cell Chem Biol*. 2016;23(10):1193-205. doi: 10.1016/j.chembiol.2016.06.017. PubMed PMID: 27693059.
130. Hayden MS, Ghosh S. Signaling to NF-kappaB. *Genes Dev*. 2004;18(18):2195-224. Epub 2004/09/17. doi: 10.1101/gad.1228704. PubMed PMID: 15371334.
131. Kasibhatla S, Brunner T, Genestier L, Echeverri F, Mahboubi A, Green DR. DNA damaging agents induce expression of Fas ligand and subsequent apoptosis in T lymphocytes via the activation of NF-kappa B and AP-1. *Mol Cell*. 1998;1(4):543-51. Epub 1998/07/14. PubMed PMID: 9660938.
132. Lawrence T. The nuclear factor NF-kappaB pathway in inflammation. *Cold Spring Harb Perspect Biol*. 2009;1(6):a001651. Epub 2010/05/12. doi: 10.1101/cshperspect.a001651. PubMed PMID: 20457564; PubMed Central PMCID: PMC2882124.

133. Pahl HL. Activators and target genes of Rel/NF-kappaB transcription factors. *Oncogene*. 1999;18(49):6853-66. Epub 1999/12/22. doi: 10.1038/sj.onc.1203239. PubMed PMID: 10602461.
134. Tak PP, Firestein GS. NF-kappaB: a key role in inflammatory diseases. *J Clin Invest*. 2001;107(1):7-11. Epub 2001/01/03. doi: 10.1172/JCI11830. PubMed PMID: 11134171; PubMed Central PMCID: PMCPMC198552.
135. Wajant H, Scheurich P. TNFR1-induced activation of the classical NF-kappaB pathway. *FEBS J*. 2011;278(6):862-76. Epub 2011/01/15. doi: 10.1111/j.1742-4658.2011.08015.x. PubMed PMID: 21232017.
136. Lewis CE, Pollard JW. Distinct role of macrophages in different tumor microenvironments. *Cancer Res*. 2006;66(2):605-12. Epub 2006/01/21. doi: 10.1158/0008-5472.CAN-05-4005. PubMed PMID: 16423985.
137. Staudt LM. Oncogenic activation of NF-kappaB. *Cold Spring Harb Perspect Biol*. 2010;2(6):a000109. Epub 2010/06/03. doi: 10.1101/cshperspect.a000109. PubMed PMID: 20516126; PubMed Central PMCID: PMCPMC2869521.
138. Marx J. Cancer research. Inflammation and cancer: the link grows stronger. *Science*. 2004;306(5698):966-8. Epub 2004/11/06. doi: 10.1126/science.306.5698.966. PubMed PMID: 15528423.
139. Schottenfeld D, Beebe-Dimmer J. Chronic inflammation: a common and important factor in the pathogenesis of neoplasia. *CA Cancer J Clin*. 2006;56(2):69-83. Epub 2006/03/04. PubMed PMID: 16514135.
140. DiDonato JA, Mercurio F, Karin M. NF-kappaB and the link between inflammation and cancer. *Immunol Rev*. 2012;246(1):379-400. Epub 2012/03/23. doi: 10.1111/j.1600-065X.2012.01099.x. PubMed PMID: 22435567.
141. Clark K, Nanda S, Cohen P. Molecular control of the NEMO family of ubiquitin-binding proteins. *Nat Rev Mol Cell Biol*. 2013;14(10):673-85. Epub 2013/08/31. doi: 10.1038/nrm3644. PubMed PMID: 23989959.
142. Haas TL, Emmerich CH, Gerlach B, Schmukle AC, Cordier SM, Rieser E, et al. Recruitment of the linear ubiquitin chain assembly complex stabilizes the TNF-R1 signaling complex and is required for TNF-mediated gene induction. *Mol Cell*. 2009;36(5):831-44. Epub 2009/12/17. doi: 10.1016/j.molcel.2009.10.013. PubMed PMID: 20005846.
143. Hayden MS, Ghosh S. Shared principles in NF-kappaB signaling. *Cell*. 2008;132(3):344-62. Epub 2008/02/13. doi: 10.1016/j.cell.2008.01.020. PubMed PMID: 18267068.
144. Hsu H, Huang J, Shu HB, Baichwal V, Goeddel DV. TNF-dependent recruitment of the protein kinase RIP to the TNF receptor-1 signaling complex. *Immunity*. 1996;4(4):387-96. Epub 1996/04/01. PubMed PMID: 8612133.
145. Kulathu Y, Akutsu M, Bremm A, Hofmann K, Komander D. Two-sided ubiquitin binding explains specificity of the TAB2 NZF domain. *Nat Struct Mol Biol*. 2009;16(12):1328-30. Epub 2009/11/26. doi: 10.1038/nsmb.1731. PubMed PMID: 19935683.

146. Ikeda F, Deribe YL, Skanland SS, Stieglitz B, Grabbe C, Franz-Wachtel M, et al. SHARPIN forms a linear ubiquitin ligase complex regulating NF-kappaB activity and apoptosis. *Nature*. 2011;471(7340):637-41. Epub 2011/04/02. doi: 10.1038/nature09814. PubMed PMID: 21455181; PubMed Central PMCID: PMC3085511.
147. Keenan AB, Jenkins SL, Jagodnik KM, Koplev S, He E, Torre D, et al. The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. *Cell Syst*. 2018;6(1):13-24. Epub 2017/12/05. doi: 10.1016/j.cels.2017.11.001. PubMed PMID: 29199020; PubMed Central PMCID: PMC5799026.
148. Lee RE, Walker SR, Savery K, Frank DA, Gaudet S. Fold change of nuclear NF-kappaB determines TNF-induced transcription in single cells. *Mol Cell*. 2014;53(6):867-79. Epub 2014/02/18. doi: 10.1016/j.molcel.2014.01.026. PubMed PMID: 24530305; PubMed Central PMCID: PMC3977799.
149. Park YC, Burkitt V, Villa AR, Tong L, Wu H. Structural basis for self-association and receptor recognition of human TRAF2. *Nature*. 1999;398(6727):533-8. Epub 1999/04/17. doi: 10.1038/19110. PubMed PMID: 10206649.
150. Park YC, Ye H, Hsia C, Segal D, Rich RL, Liou HC, et al. A novel mechanism of TRAF signaling revealed by structural and functional analyses of the TRADD-TRAF2 interaction. *Cell*. 2000;101(7):777-87. Epub 2000/07/13. PubMed PMID: 10892748.
151. Tarantino N, Tinevez JY, Crowell EF, Boisson B, Henriques R, Mhlanga M, et al. TNF and IL-1 exhibit distinct ubiquitin requirements for inducing NEMO-IKK supramolecular structures. *J Cell Biol*. 2014;204(2):231-45. Epub 2014/01/22. doi: 10.1083/jcb.201307172. PubMed PMID: 24446482; PubMed Central PMCID: PMC3897181.
152. Zhang Q, Gupta S, Schipper DL, Kowalczyk GJ, Mancini AE, Faeder JR, et al. NF-kappaB Dynamics Discriminate between TNF Doses in Single Cells. *Cell Syst*. 2017;5(6):638-45 e5. Epub 2017/11/13. doi: 10.1016/j.cels.2017.10.011. PubMed PMID: 29128333; PubMed Central PMCID: PMC5746429.
153. Wong VC, Bass VL, Bullock ME, Chavali AK, Lee REC, Mothes W, et al. NF-kappaB-Chromatin Interactions Drive Diverse Phenotypes by Modulating Transcriptional Noise. *Cell Rep*. 2018;22(3):585-99. Epub 2018/01/19. doi: 10.1016/j.celrep.2017.12.080. PubMed PMID: 29346759; PubMed Central PMCID: PMC5812697.
154. Zarnegar BJ, Wang Y, Mahoney DJ, Dempsey PW, Cheung HH, He J, et al. Noncanonical NF-kappaB activation requires coordinated assembly of a regulatory complex of the adaptors cIAP1, cIAP2, TRAF2 and TRAF3 and the kinase NIK. *Nat Immunol*. 2008;9(12):1371-8. Epub 2008/11/11. doi: 10.1038/ni.1676. PubMed PMID: 18997794; PubMed Central PMCID: PMC2676931.
155. Behar M, Barken D, Werner SL, Hoffmann A. The dynamics of signaling as a pharmacological target. *Cell*. 2013;155(2):448-61. Epub 2013/10/15. doi: 10.1016/j.cell.2013.09.018. PubMed PMID: 24120141; PubMed Central PMCID: PMC3856316.
156. Pabon NA, Xia Y, Estabrooks SK, Ye Z, Herbrand AK, Sub E, et al. Predicting protein targets for drug-like compounds using transcriptomics. *bioRxiv*. 2018.

157. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res.* 2000;28(1):235-42. Epub 1999/12/11. PubMed PMID: 10592235; PubMed Central PMCID: PMCPMC102472.
158. Narayan VA, Mohwinckel M, Pisano G, Yang M, Manji HK. Beyond magic bullets: true innovation in health care. *Nat Rev Drug Discov.* 2013;12(2):85-6. Epub 2013/02/02. doi: 10.1038/nrd3944. PubMed PMID: 23370233.
159. Hopkins AL. Network pharmacology. *Nat Biotechnol.* 2007;25(10):1110-1. Epub 2007/10/09. doi: 10.1038/nbt1007-1110. PubMed PMID: 17921993.
160. Kitano H. A robustness-based approach to systems-oriented drug design. *Nat Rev Drug Discov.* 2007;6(3):202-10. Epub 2007/02/24. doi: 10.1038/nrd2195. PubMed PMID: 17318209.
161. Ziegler S, Pries V, Hedberg C, Waldmann H. Target identification for small bioactive molecules: finding the needle in the haystack. *Angew Chem Int Ed Engl.* 2013;52(10):2744-92. Epub 2013/02/19. doi: 10.1002/anie.201208749. PubMed PMID: 23418026.
162. Merino A, Bronowska AK, Jackson DB, Cahill DJ. Drug profiling: knowing where it hits. *Drug Discov Today.* 2010;15(17-18):749-56. Epub 2010/07/06. doi: 10.1016/j.drudis.2010.06.006. PubMed PMID: 20601095.
163. Xavier Jalenacs JM. On the origins of drug polypharmacology. *Med Chem Commun.* 2012;4(1):80-7.
164. Azzaoui K, Hamon J, Faller B, Whitebread S, Jacoby E, Bender A, et al. Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem.* 2007;2(6):874-80. Epub 2007/05/12. doi: 10.1002/cmdc.200700036. PubMed PMID: 17492703.
165. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, et al. Large-scale prediction and testing of drug activity on side-effect targets. *Nature.* 2012;486(7403):361-7. Epub 2012/06/23. doi: 10.1038/nature11159. PubMed PMID: 22722194; PubMed Central PMCID: PMCPMC3383642.
166. Luu JK, Chappelov AV, McCulley TJ, Marmor MF. Acute effects of sildenafil on the electroretinogram and multifocal electroretinogram. *Am J Ophthalmol.* 2001;132(3):388-94. Epub 2001/09/01. PubMed PMID: 11530053.
167. Rothman RB, Baumann MH. Serotonergic drugs and valvular heart disease. *Expert Opin Drug Saf.* 2009;8(3):317-29. Epub 2009/06/10. doi: 10.1517/14740330902931524. PubMed PMID: 19505264; PubMed Central PMCID: PMCPMC2695569.
168. Anighoro A, Bajorath J, Rastelli G. Polypharmacology: challenges and opportunities in drug discovery. *J Med Chem.* 2014;57(19):7874-87. Epub 2014/06/20. doi: 10.1021/jm5006463. PubMed PMID: 24946140.
169. Bottegoni G, Favia AD, Recanatini M, Cavalli A. The role of fragment-based and computational methods in polypharmacology. *Drug Discov Today.* 2012;17(1-2):23-34. Epub 2011/08/26. doi: 10.1016/j.drudis.2011.08.002. PubMed PMID: 21864710.

170. Medina-Franco JL, Giulianotti MA, Welmaker GS, Houghten RA. Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discov Today*. 2013;18(9-10):495-501. Epub 2013/01/24. doi: 10.1016/j.drudis.2013.01.008. PubMed PMID: 23340113; PubMed Central PMCID: PMC3642214.
171. Burbaum J, Tobal GM. Proteomics in drug discovery. *Curr Opin Chem Biol*. 2002;6(4):427-33. Epub 2002/07/23. PubMed PMID: 12133716.
172. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br J Pharmacol*. 2007;152(1):9-20. Epub 2007/06/06. doi: 10.1038/sj.bjp.0707305. PubMed PMID: 17549047; PubMed Central PMCID: PMC1978274.
173. Jenwitheesuk E, Horst JA, Rivas KL, Van Voorhis WC, Samudrala R. Novel paradigms for drug discovery: computational multitarget screening. *Trends Pharmacol Sci*. 2008;29(2):62-71. Epub 2008/01/15. doi: 10.1016/j.tips.2007.11.007. PubMed PMID: 18190973; PubMed Central PMCID: PMC4551513.
174. Gong J, Cai C, Liu X, Ku X, Jiang H, Gao D, et al. ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. *Bioinformatics*. 2013;29(14):1827-9. doi: 10.1093/bioinformatics/btt270. PubMed PMID: 23712658.
175. Wang L, Ma C, Wipf P, Liu H, Su W, Xie XQ. TargetHunter: an in silico target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database. *The AAPS journal*. 2013;15(2):395-406. doi: 10.1208/s12248-012-9449-z. PubMed PMID: 23292636; PubMed Central PMCID: PMC3675739.
176. Cobanoglu MC, Oltvai ZN, Taylor DL, Bahar I. BalestraWeb: efficient online evaluation of drug-target interactions. *Bioinformatics*. 2015;31(1):131-3. doi: 10.1093/bioinformatics/btu599. PubMed PMID: 25192741; PubMed Central PMCID: PMC4271144.
177. Chen YZ, Zhi DG. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins*. 2001;43(2):217-26. Epub 2001/03/29. PubMed PMID: 11276090.
178. Kharkar PS, Warriar S, Gaud RS. Reverse docking: a powerful tool for drug repositioning and drug rescue. *Future Med Chem*. 2014;6(3):333-42. doi: 10.4155/fmc.13.207. PubMed PMID: 24575968.
179. Chen X, Ung CY, Chen Y. Can an in silico drug-target search method be used to probe potential mechanisms of medicinal plant ingredients? *Nat Prod Rep*. 2003;20(4):432-44. Epub 2003/09/11. PubMed PMID: 12964838.
180. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. 2010;31(2):455-61. Epub 2009/06/06. doi: 10.1002/jcc.21334. PubMed PMID: 19499576; PubMed Central PMCID: PMC3041641.
181. Brenke R, Kozakov D, Chuang GY, Beglov D, Hall D, Landon MR, et al. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques.

- Bioinformatics. 2009;25(5):621-7. Epub 2009/01/30. doi: 10.1093/bioinformatics/btp036. PubMed PMID: 19176554; PubMed Central PMCID: PMCPMC2647826.
182. Wingert BM, Oerlemans R, Camacho CJ. Optimal affinity ranking for automated virtual screening validated in prospective D3R grand challenges. *J Comput Aided Mol Des.* 2018;32(1):287-97. Epub 2017/09/18. doi: 10.1007/s10822-017-0065-y. PubMed PMID: 28918599; PubMed Central PMCID: PMCPMC5771500.
 183. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *J Cheminform.* 2011;3:33. Epub 2011/10/11. doi: 10.1186/1758-2946-3-33. PubMed PMID: 21982300; PubMed Central PMCID: PMCPMC3198950.
 184. Rego N, Koes D. 3Dmol.js: molecular visualization with WebGL. *Bioinformatics.* 2015;31(8):1322-4. Epub 2014/12/17. doi: 10.1093/bioinformatics/btu829. PubMed PMID: 25505090; PubMed Central PMCID: PMCPMC4393526.
 185. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 2018;46(D1):D1074-D82. Epub 2017/11/11. doi: 10.1093/nar/gkx1037. PubMed PMID: 29126136; PubMed Central PMCID: PMCPMC5753335.
 186. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. *Nucleic Acids Res.* 2017;45(D1):D945-D54. Epub 2016/12/03. doi: 10.1093/nar/gkw1074. PubMed PMID: 27899562; PubMed Central PMCID: PMCPMC5210557.
 187. Li Z, Han P, You ZH, Li X, Zhang Y, Yu H, et al. In silico prediction of drug-target interaction networks based on drug chemical structure and protein sequences. *Sci Rep.* 2017;7(1):11174. Epub 2017/09/13. doi: 10.1038/s41598-017-10724-0. PubMed PMID: 28894115; PubMed Central PMCID: PMCPMC5593914.
 188. Eric S, Ke S, Barata T, Solmajer T, Antic Stankovic J, Juranic Z, et al. Target fishing and docking studies of the novel derivatives of aryl-aminopyridines with potential anticancer activity. *Bioorg Med Chem.* 2012;20(17):5220-8. Epub 2012/07/31. doi: 10.1016/j.bmc.2012.06.051. PubMed PMID: 22841617.
 189. Luo Q, Zhao L, Hu J, Jin H, Liu Z, Zhang L. The scoring bias in reverse docking and the score normalization strategy to improve success rate of target fishing. *PLoS One.* 2017;12(2):e0171433. Epub 2017/02/15. doi: 10.1371/journal.pone.0171433. PubMed PMID: 28196116; PubMed Central PMCID: PMCPMC5308821.